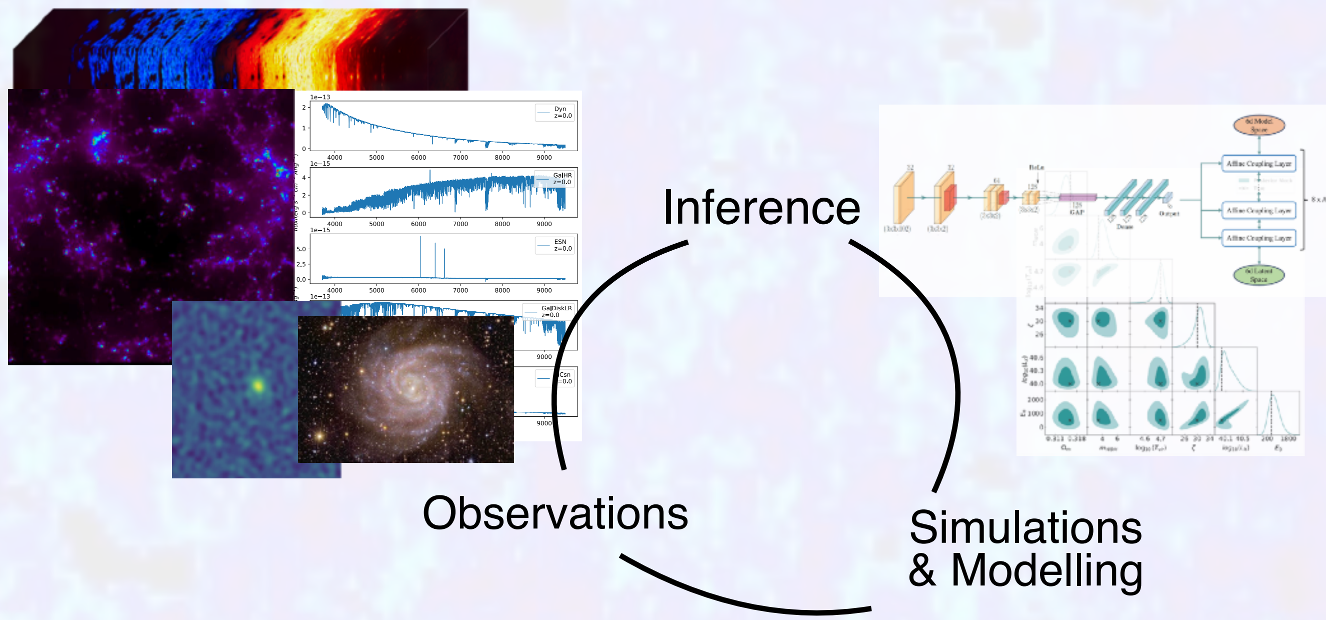
Machine Learning for Astrophysics & Cosmology



Caroline Heneka, group leader, ITP Heidelberg

‘Computer Vision Astrophysics and Cosmology’

Physics in the AI era, Pisa, September 25th 2024

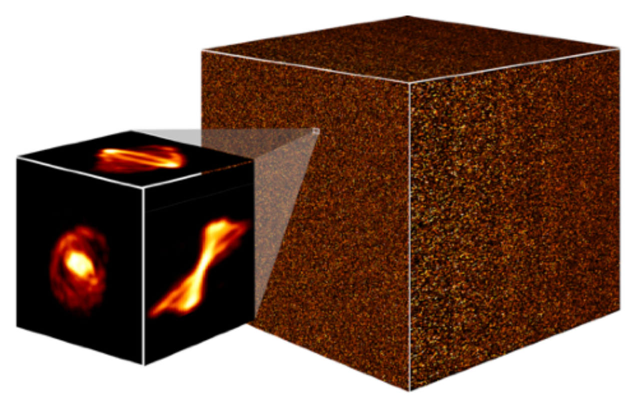
Cosmology & ML group @ ITP

Our goal
Learn about cosmology, large-scale structure, the high-redshift Universe (Reionization) and develop the suitable modern ML toolkit.

About myself:
B.Sc. and M.Sc. in Physics (Heidelberg)
PhD in Physics 2017 (Copenhagen)
Postdoc: SNS Pisa, UHH Hamburg, + DLR
Since 10/22 Group Leader

Our research:
- Computational astrophysics / cosmology
- Intensity mapping
- Large (radio) surveys, SKA & LOFAR

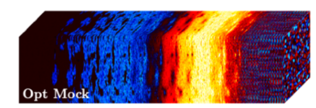
Specifically the modern ML toolkit for cosmology and large-scale surveys:
- Emulation, generation
- Inference
- Classification, anomaly detection
- Computer Vision tasks in astronomy



STRUCTURES CLUSTER OF EXCELLENCE
Daimler und Benz Stiftung
VolkswagenStiftung FREIGEIST FELLOWSHIP DER VOLKSWAGENSTIFTUNG

How will Astrophysics and Cosmology advance in coming years?

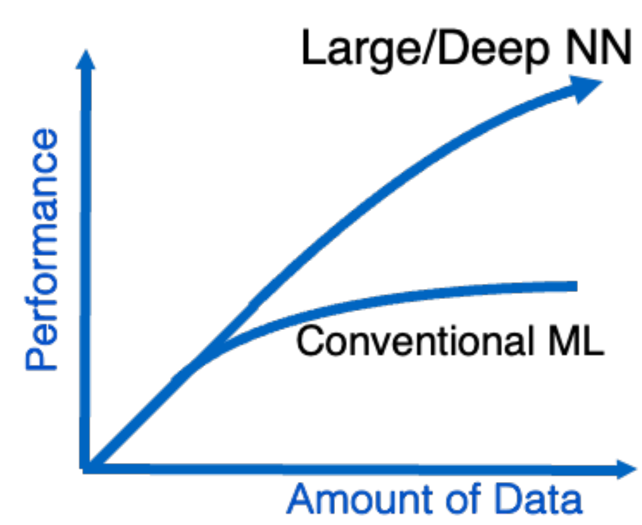
Understanding



Survey Data

Data Science
'ML + AI'

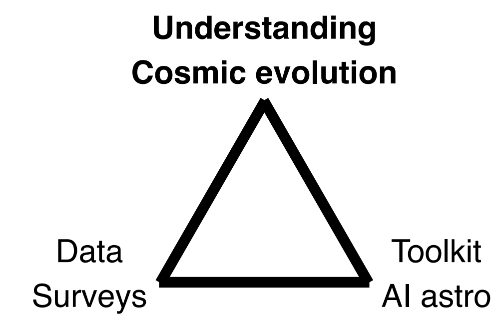
Credit: SKAO

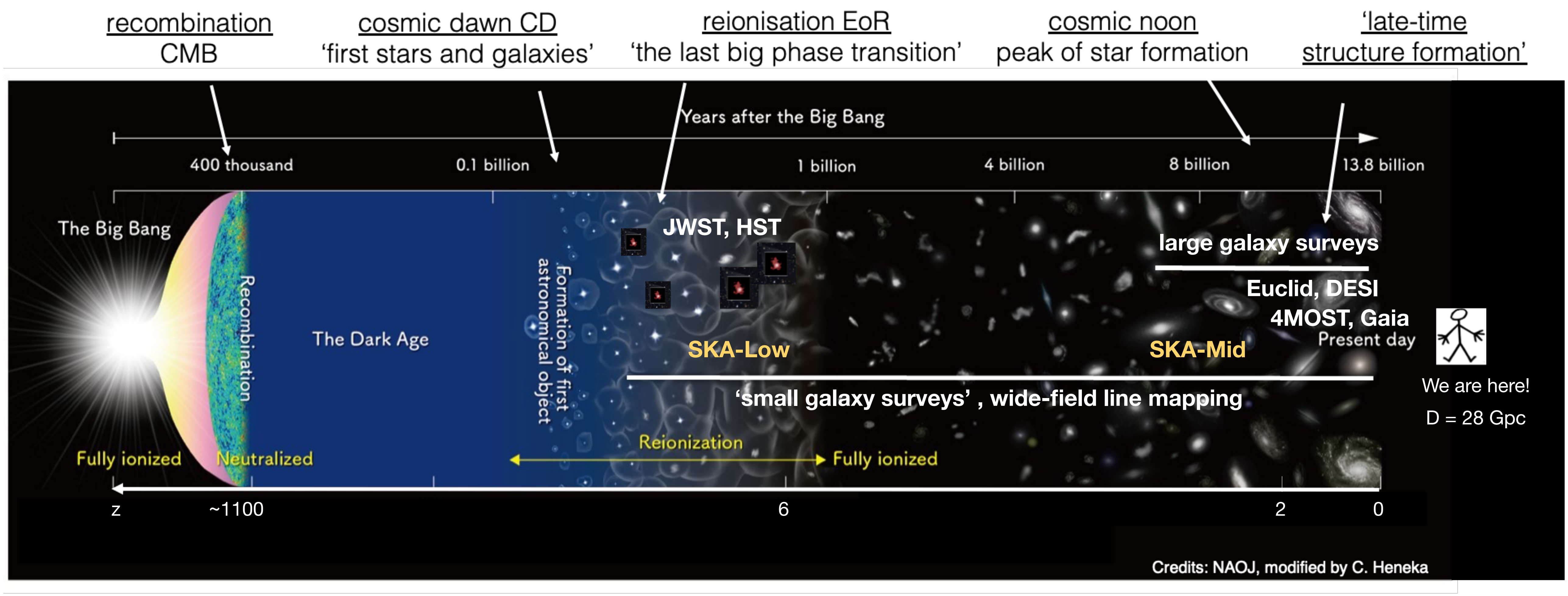


Efficient data reduction
Automation

Extract more & less biased information
Data mining

Where we stand: Data revolution and cosmic evolution





Our goal:

Learn about astrophysical & cosmological evolution
across cosmic time and scales

Coming decade: push to map up to **80% of the observable Universe**

… what does 80% of the observable Universe mean?

Modelling challenges

True LSS probes $\longrightarrow$ orders of magnitude of scales up to the ultra-large

…what does 80% of the observable Universe even mean?

APOD, NASA, License & Credit: Wikipedia, Pablo Carlos Budassi

Observable Universe:

d ~ 28 Gpc (x3 Glyr)

80% if this:

d ~ 22 Gpc

Let's say we resolve (only) ~Mpc

$\longrightarrow$ about 3-4 orders of magnitude

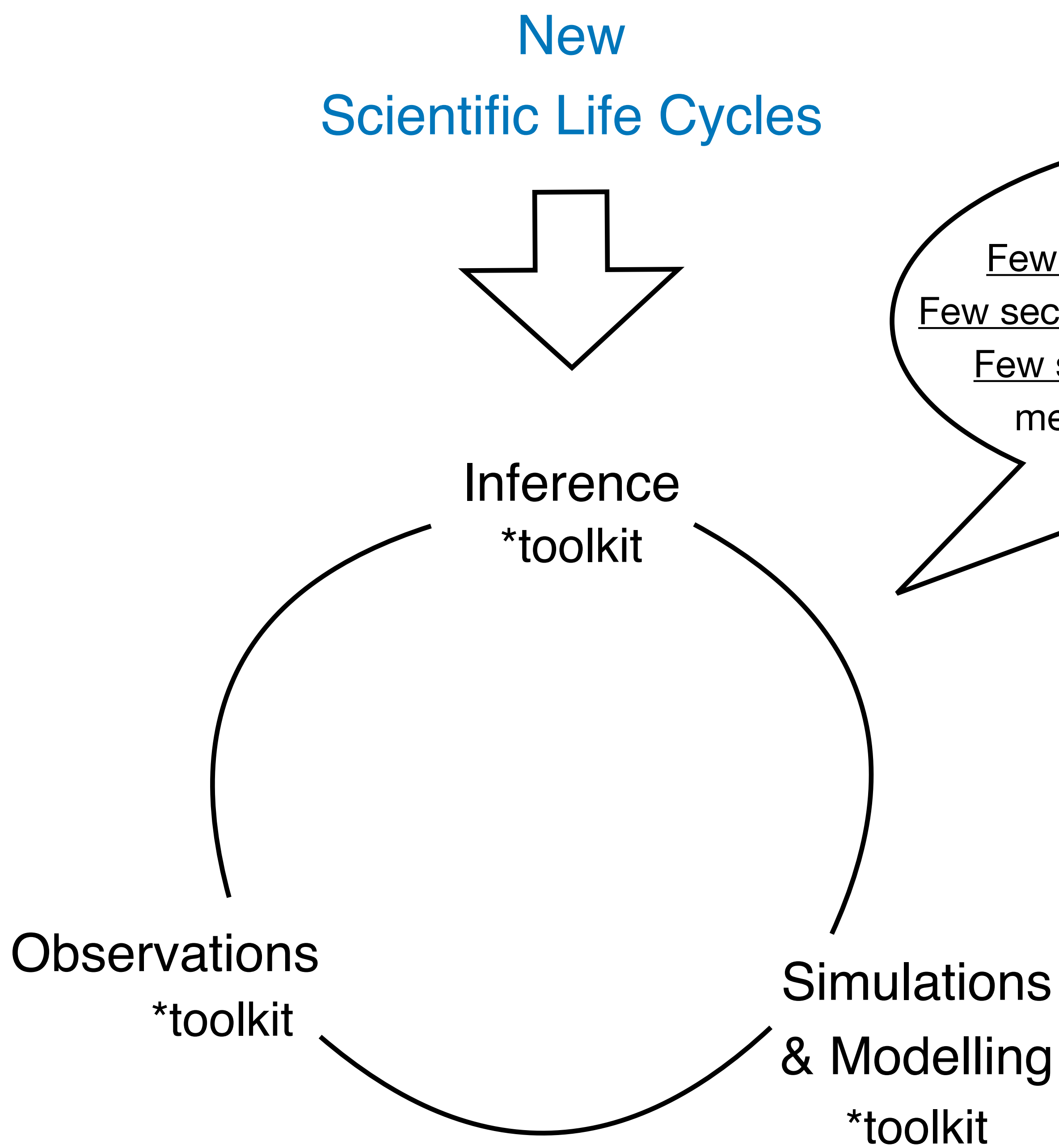
$\longrightarrow$ about 10^9-10^{10} modes!

… at some point we sub-grid model and/or change modelling approach

Understanding
Cosmic evolution

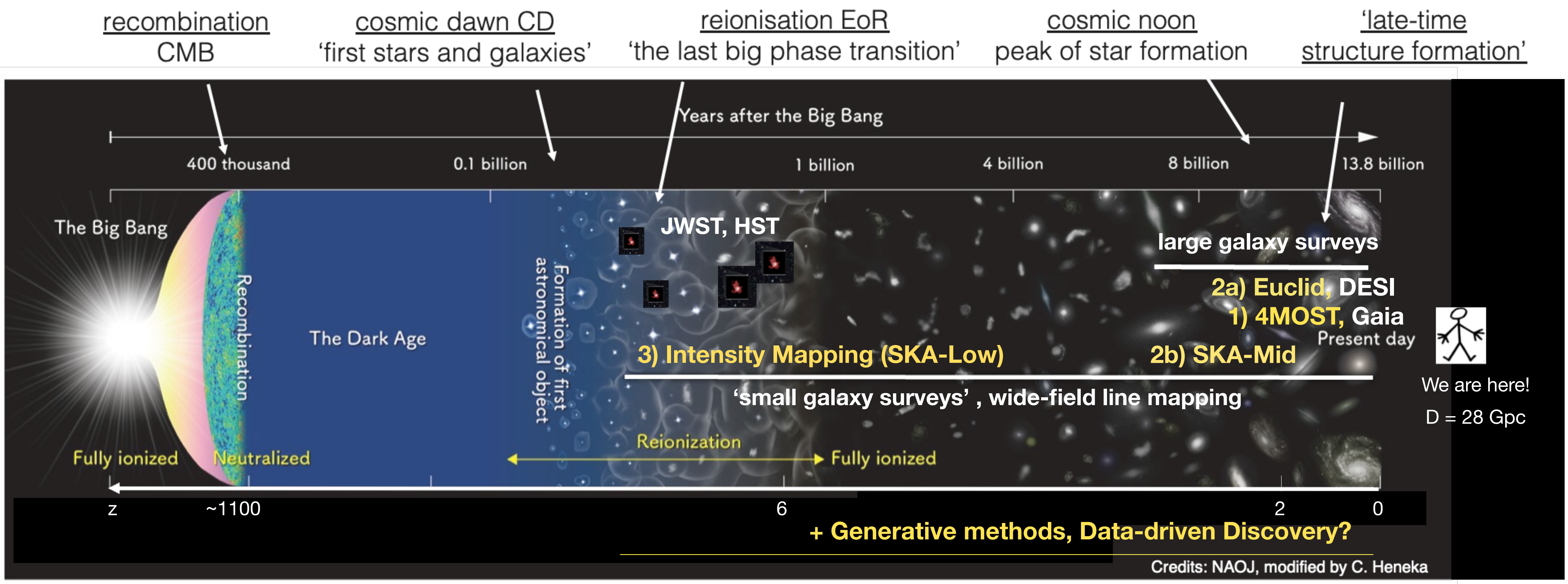
How will Astrophysics and Cosmology advance in coming years?

We need a versatile ML/AI toolkit*.



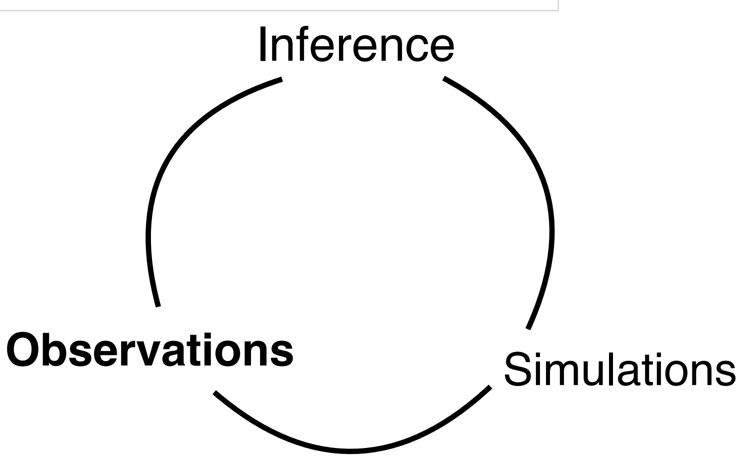
Examples:
Few sec: Classification 40.000 spectra
Few sec: 7-parameter inference ~100MB cube
Few sec: detection, segmentation & flux measurement O(100-1000) sources

Research Highlights: Astronomical Data Science and Artificial Intelligence



Select Highlights

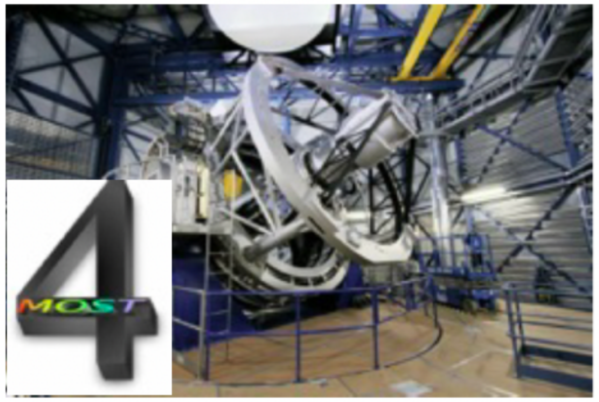
1) **Classification**
2) Source detection & characterisation
3) Simulation-based inference
+ Generative methods, Data-driven Discovery



1) Classification and triggering for large astronomical surveys

4MOST: On-the-fly classification of spectra (1D)

- 5-year survey
- wide-field, fibre-fed, optical spectroscopy
- on ESO’s 4-m-class telescope VISTA
- 2.5-degree diameter field-of-view, 2436 fibres
- HRS R ≈ 18000 – 21000, LRS R ≈ 4000 – 7500
- 20mio. (LRS), 3mio. (HRS) sources

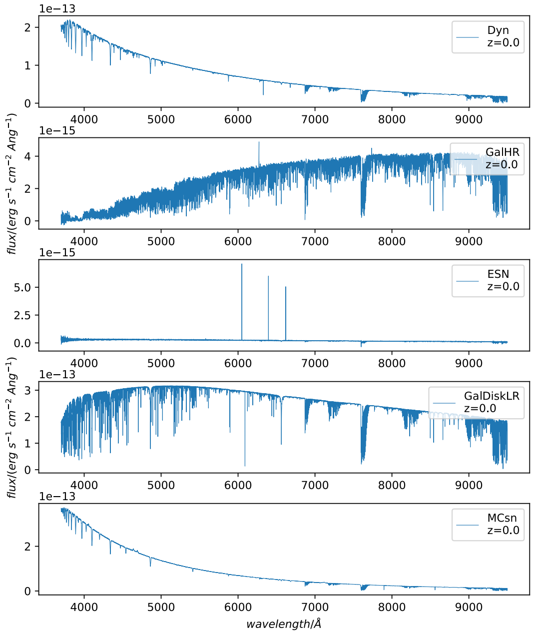


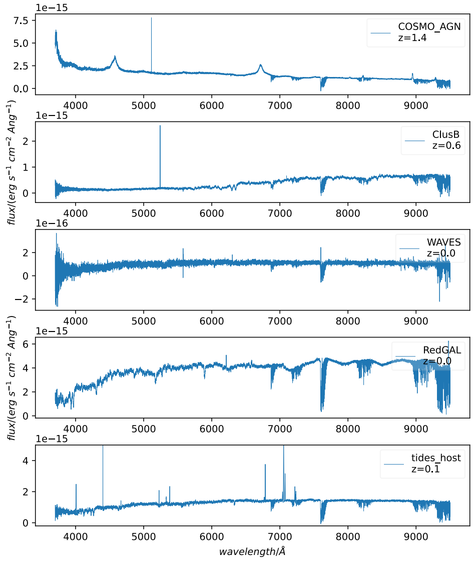
https://www.4most.eu Credit: ESO

Goal: Data-driven classification pipeline layer (galactic & extragalactic sources)

Classification infrastructure working group, led by: N. Napolitano & C. Heneka

→ *Benchmark with SDSS archival spectra:*

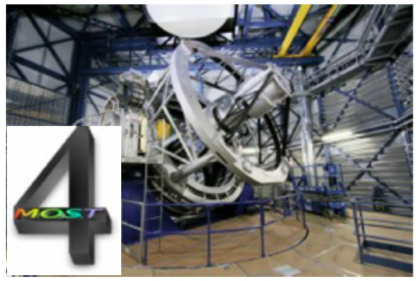




@SDSS

1) Classification and triggering for large astronomical surveys

4MOST: On-the-fly classification of spectra (1D)



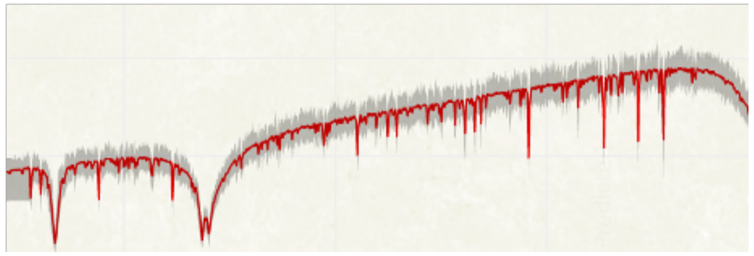
<u>Goal:</u> Data-driven classification pipeline layer (galactic & extragalactic sources)

Classification infrastructure working group, led by: N. Napolitano & C. Heneka

→ Probabilistic multi-classifier

<u>For class:</u>
Convolutional network variants

<u>For class uncertainties:</u>
Bayesian neural networks and contrastive learning



++ competitive with template fitting

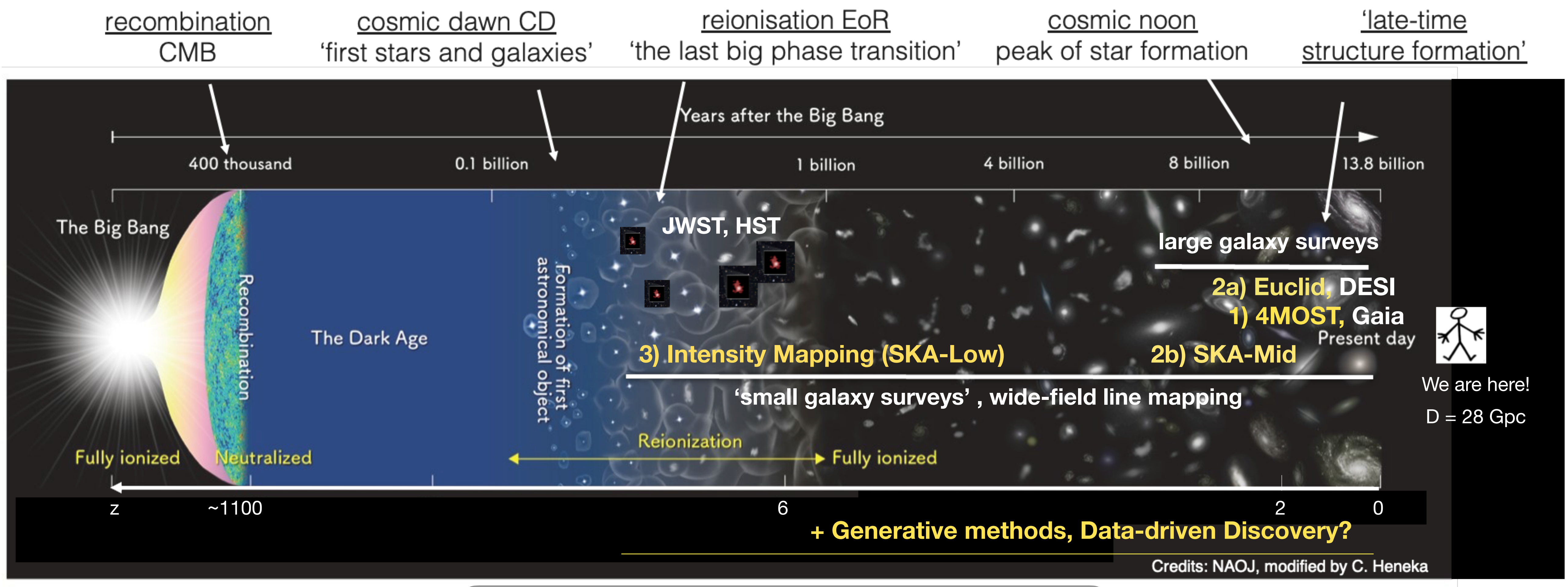
Examples:
<u>Few sec:</u> Classification 40.000 spectra
<u>Few sec:</u> 7-parameter inference ~100MB cube
<u>Few sec:</u> detection, segmentation & flux measurement O(100-1000) sources

Zhong, Napolitano, Heneka+ arXiv:2311.04146

	GALAXY_AGN	GALAXY_STARBURST	GALAXY_STARFORMING	GALAXY_nan	QSO_BROADLINE	QSO_nan	STAR_A0	STAR_F5	STAR_F9	STAR_G2	STAR_K1	STAR_K3	STAR_K5
STAR_K5	1 0.00			3 0.00		1 0.00			1 0.00		1 0.00	142 0.05	2920 0.97
STAR_K3						2 0.00			8 0.00		186 0.06	2719 0.91	75 0.03
STAR_K1			1 0.00			3 0.00			126 0.04		2748 0.92	129 0.04	
STAR_G2			1 0.00	1 0.00		3 0.00		14 0.00	139 0.05	2882 0.96			
STAR_F9				3 0.00					2725 0.91	95 0.03	64 0.02	9 0.00	2 0.00
STAR_F5		1 0.00				1 0.00	165 0.06	2899 0.97		21 0.01			
STAR_A0						2 0.00	2832 0.94	87 0.03			1 0.00		1 0.00
QSO_nan	9 0.00	6 0.00	4 0.00	32 0.01	284 0.09	2716 0.91	2 0.00			2 0.00			1 0.00
QSO_BROADLINE	3 0.00		1 0.00		2703 0.90	187 0.06							
GALAXY_nan	103 0.03	19 0.01	100 0.03	2716 0.91	2 0.00	65 0.02	1 0.00		1 0.00			1 0.00	1 0.00
GALAXY_STARFORMING	87 0.03	163 0.05	2608 0.87	126 0.04		3 0.00							
GALAXY_STARBURST	17 0.01	2794 0.93	135 0.04	21 0.01									
GALAXY_AGN	2780 0.93	17 0.01	150 0.05	98 0.03	11 0.00	17 0.01							

Actual

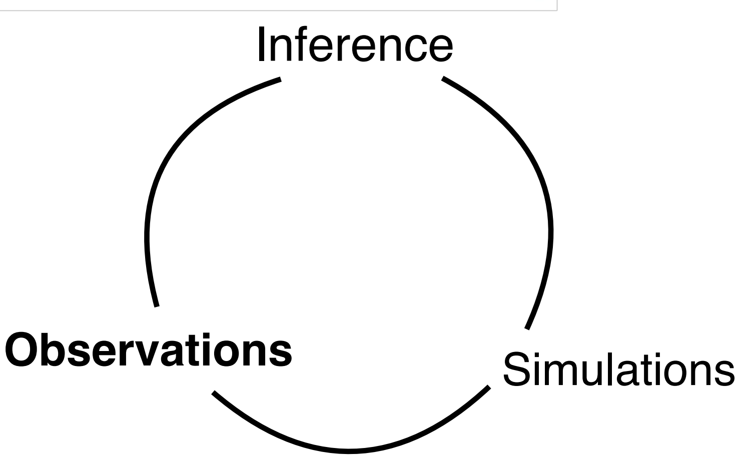
Research Highlights: Astronomical Data Science and Artificial Intelligence



Select Highlights

1) Classification
2) **Source detection & characterisation**
3) Simulation-based inference

\+ Generative methods, Data-driven Discovery



2a) The deblending problem

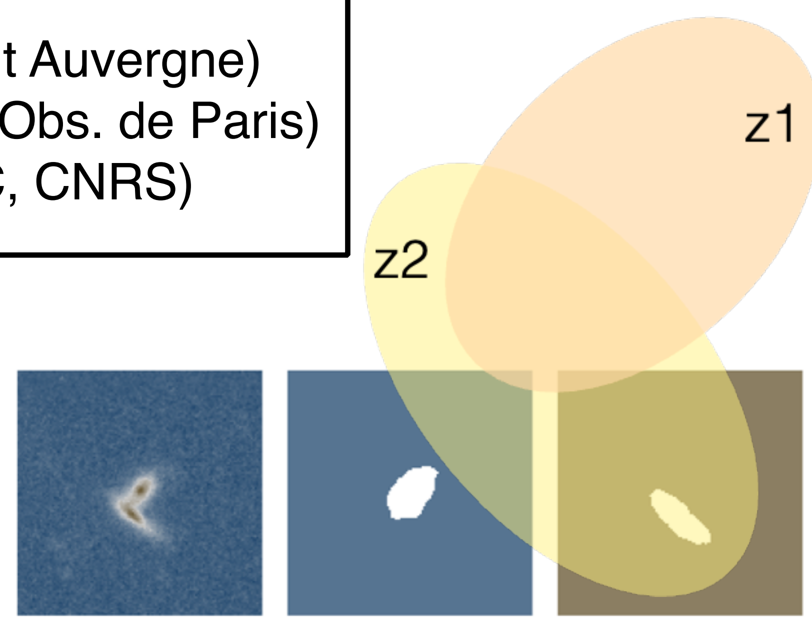
Example: Optical source detection & characterisation

Goal: ‘Good’ photometry for surveys with high blended fraction
- avoid bias!

Challenge: Galaxies are ‘transparent’

COIN network:
Emille Ishida (U. Clermont Auvergne)
Marc Huertas-Company (Obs. de Paris)
Alexandre Boucaud (APC, CNRS)

coindeblend



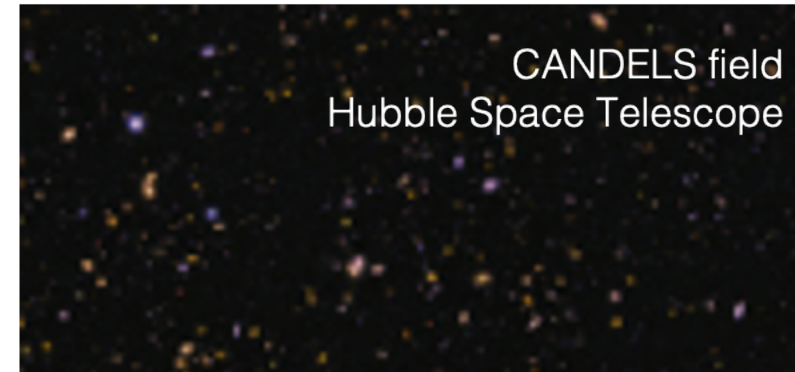
Boucaud, Huertas-Company, Heneka+ 20, arXiv:1905.01324

Lily Hu+ 2017

Similar challenge: Overlapping chromosomes

2a) The deblending problem

Example: Optical source detection & characterisation

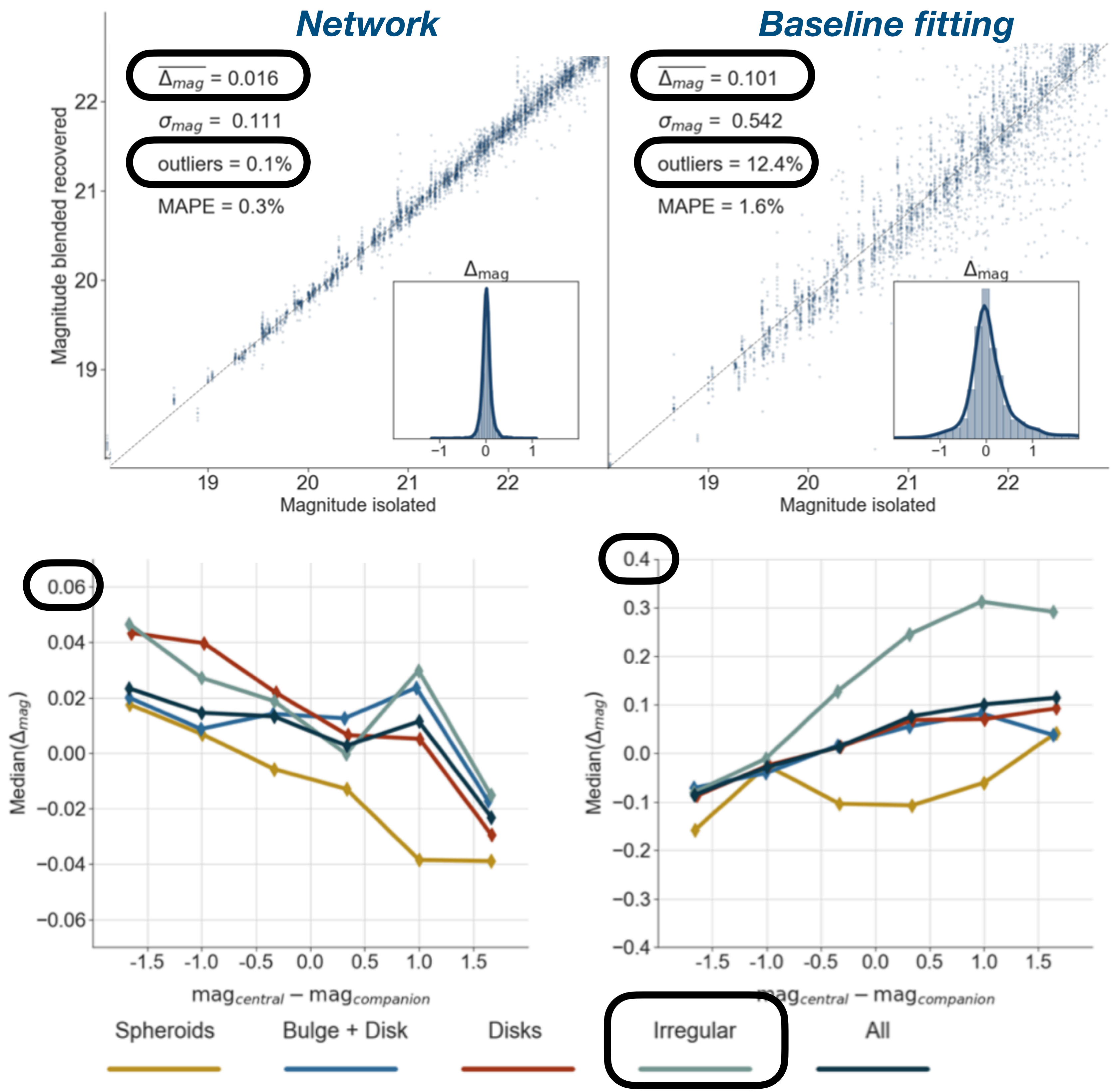


Goal: ‘Good’ photometry for surveys with high blended fraction
- avoid bias!

Galaxy morphology

Credit: Euclid, ESA

2a) Optical source detection and characterisation



Precise Photometry

~~Bias~~

Irregular

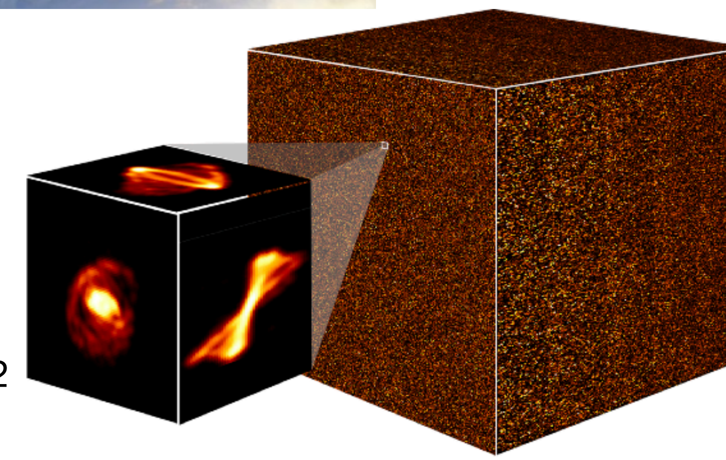
coindeblend

Boucaud, Huertas-Company, Heneka+ 20

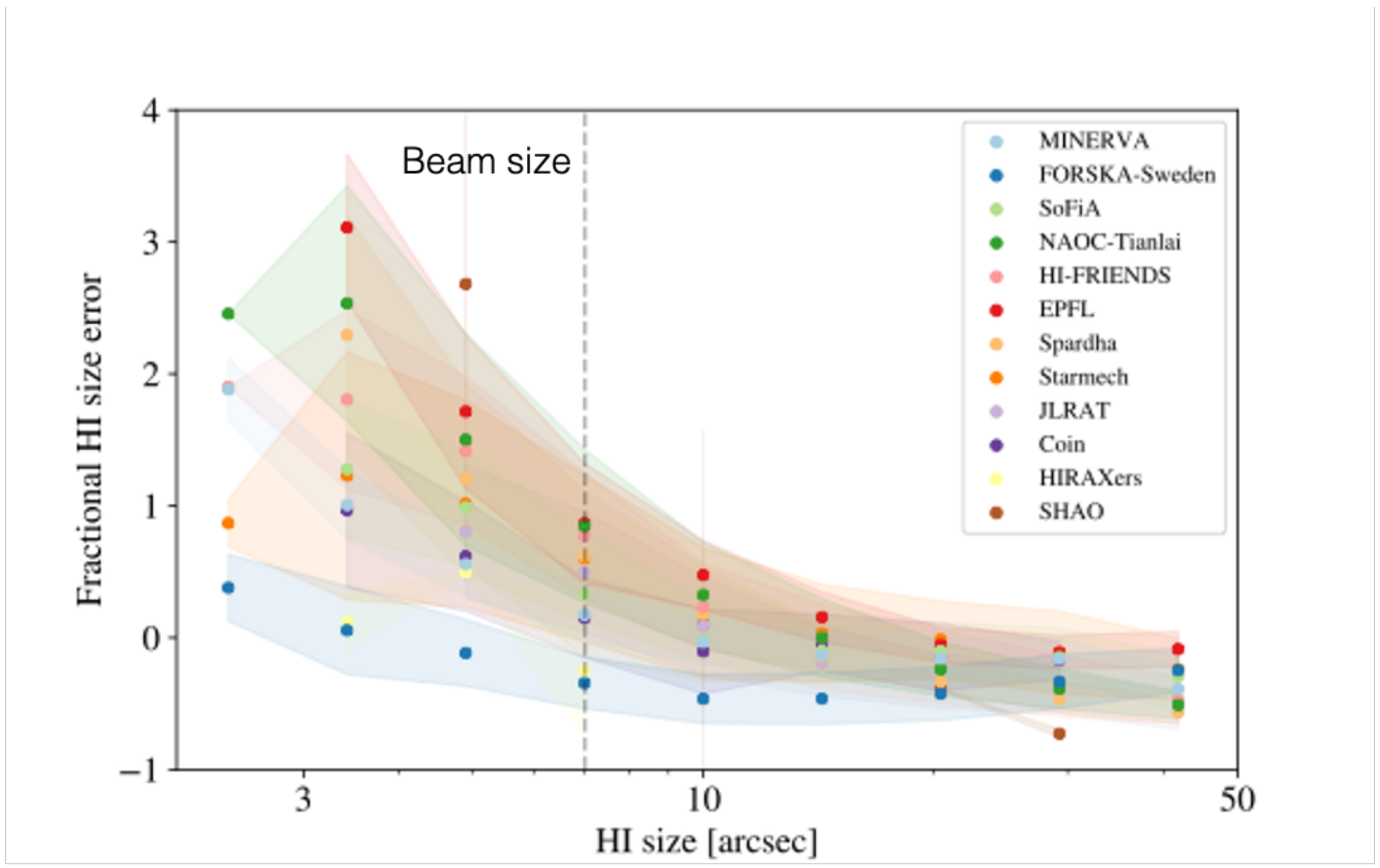
2b) Radio source detection and characterisation

Example: Source detection in tomographic data

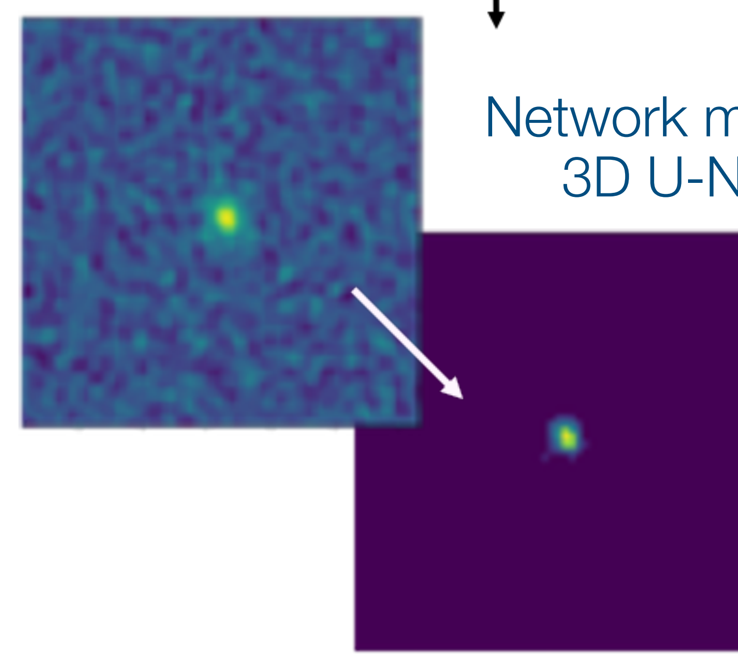
Credit: SKAO, SDC2



- 3D better than stitching of 2D + 1D
- High-fidelity 3D reconstructions
- **Good prior for characterisation tasks via nets:**



Total dimensions: (25,714 x 25,714 x 6,667) vox

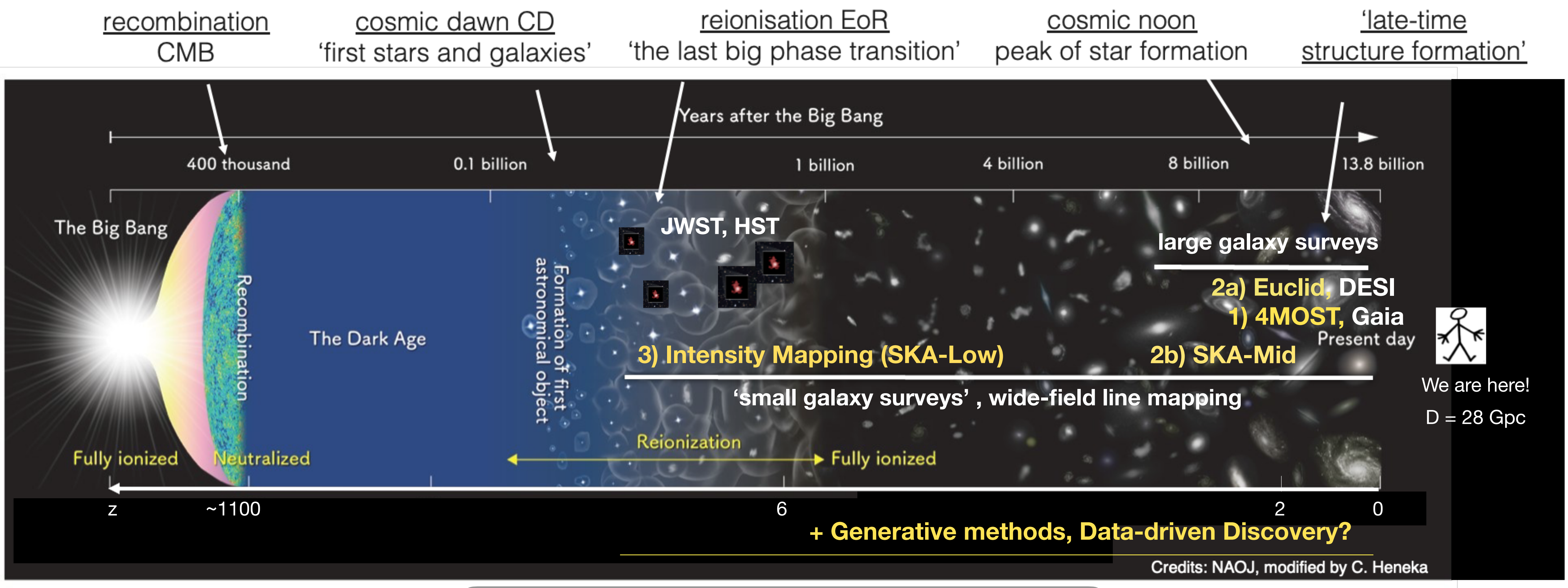


Network model:
3D U-Net

@HPC/GPU Jean Zay (Idris)

Hartley+ 23 (incl. Heneka), arXiv:2303.07943
Heneka 23, arXiv:2311.17553

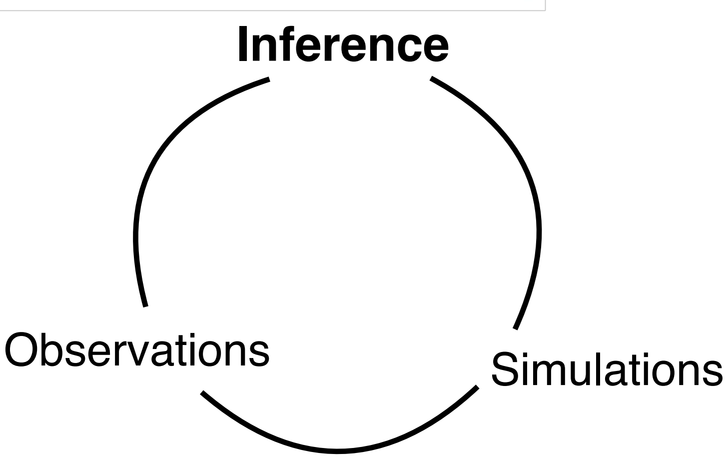
Research Highlights: Astronomical Data Science and Artificial Intelligence



Select Highlights

1) Classification
2) Source detection & characterisation
3) **Simulation-based inference**

\+ Generative methods, Data-driven Discovery



3) 3D Simulation-based inference (SBI) for the SKA

SKA: TB/s, few EB/day
Archive: ~700 PB/yr

Why care?

Tomography of >50% of the Universe

True 'Big Data'

non-linear, non-Gaussian signal

→ MCMC becomes slow and biased

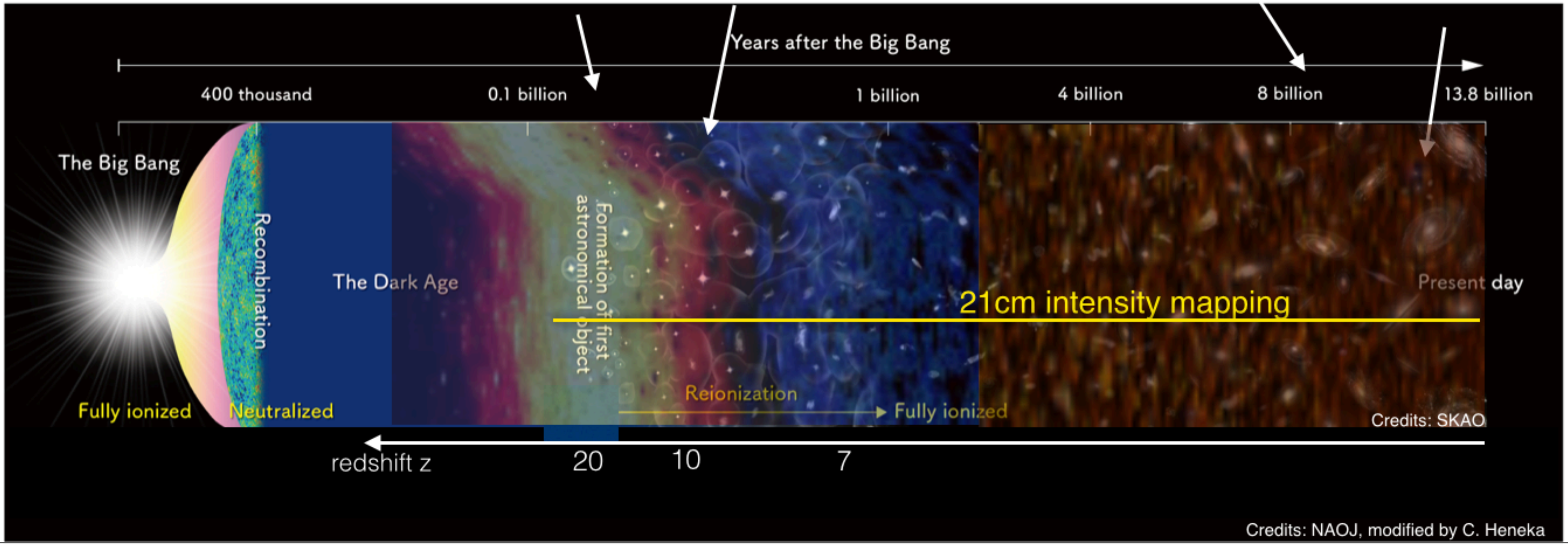
→ Move to full likelihood(-free) inference with networks

cosmic dawn **CD**
'first stars and galaxies'

reionisation **EoR**
'the last big phase transition'

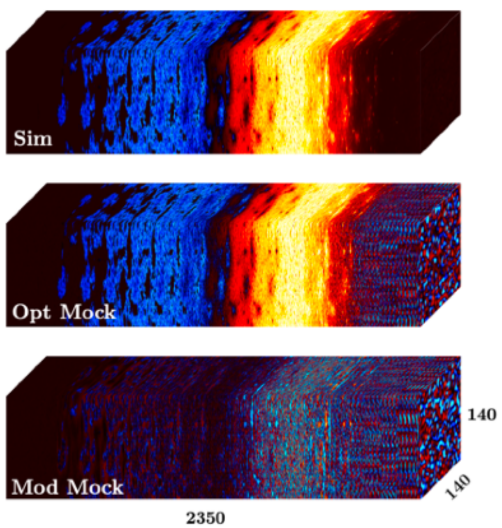
peak of
star formation

'late-time
structure formation'

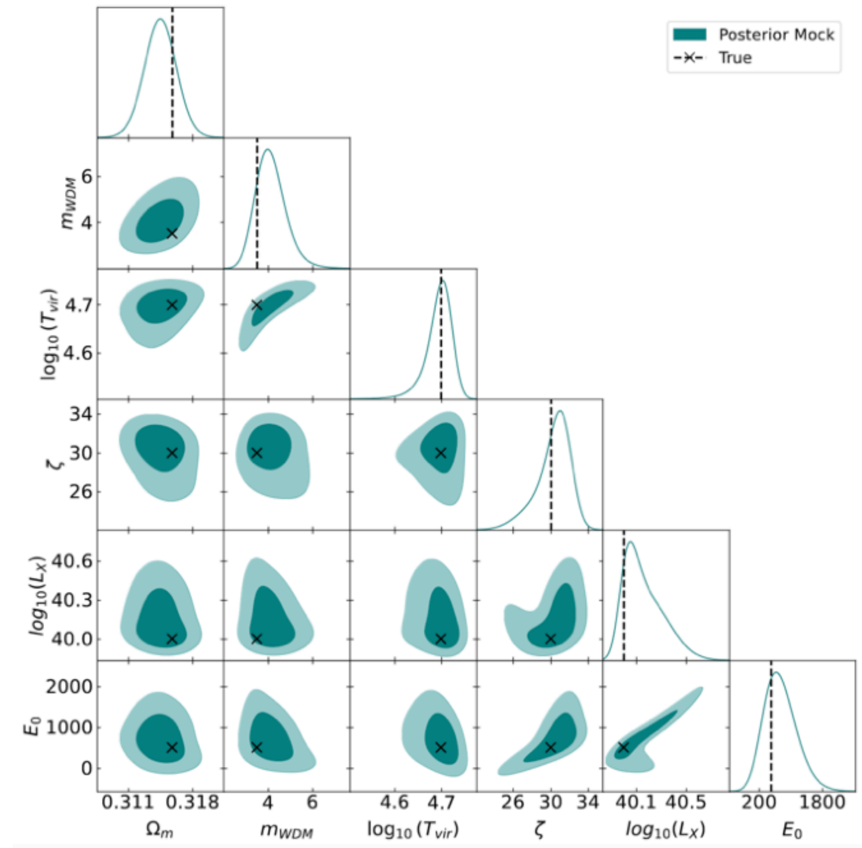


We are here

3) Simulation-based inference (SBI) for the SKA



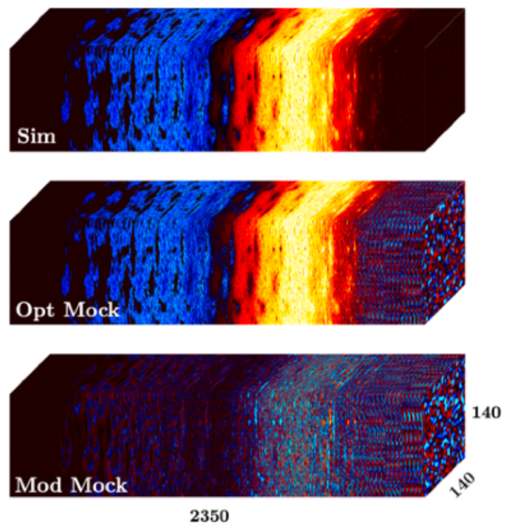
?



Neutsch, Heneka, Brüggen (2022), arXiv:2201.07587

Schosser, Heneka, Plehn, arXiv:2401.04174

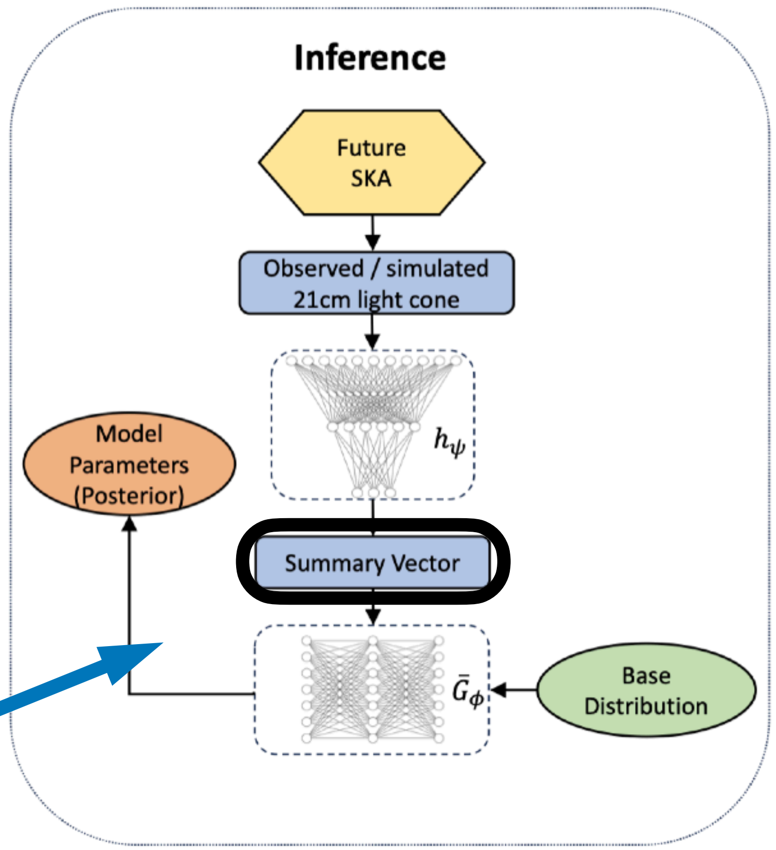
3) Simulation-based inference (SBI) for the SKA

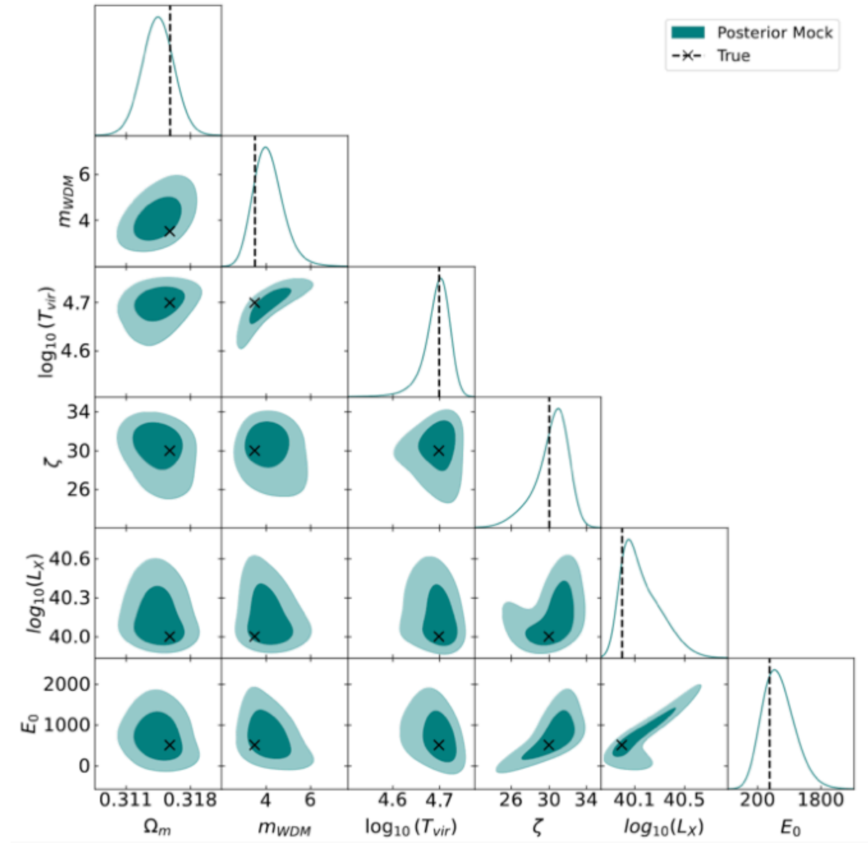


SBI with flows/cINN

based on BayesFlow

arXiv:2003.06281





‘The cool’ here:

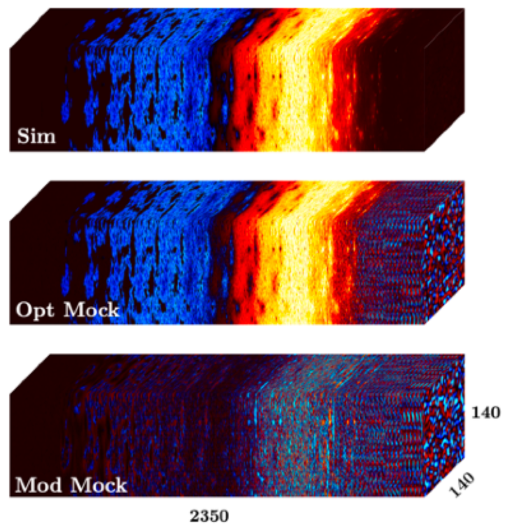
End-to-end training

Learned optimal summary

Neutsch, Heneka, Brüggen (2022), arXiv:2201.07587

Schosser, Heneka, Plehn, arXiv:2401.04174

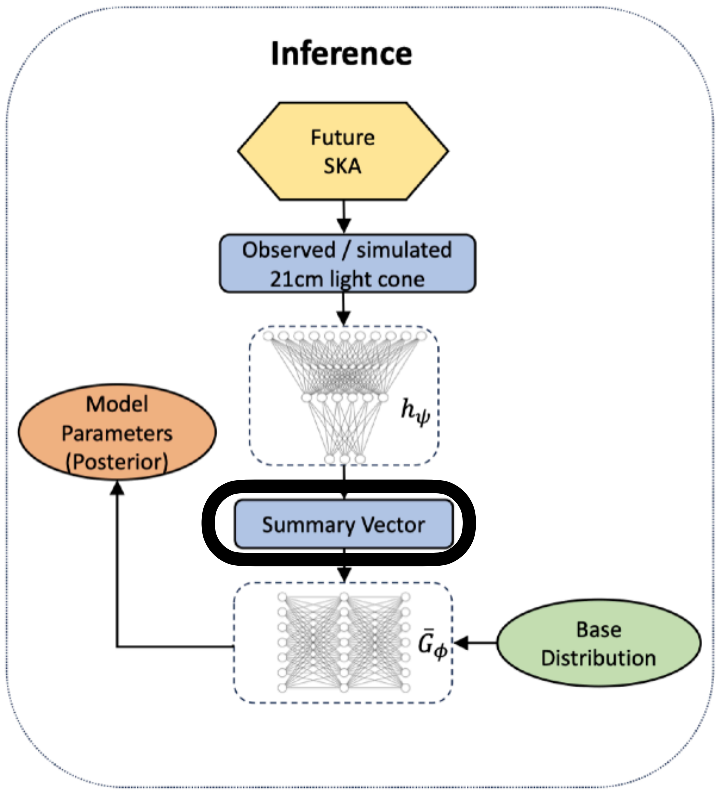
3) Simulation-based inference (SBI) for the SKA

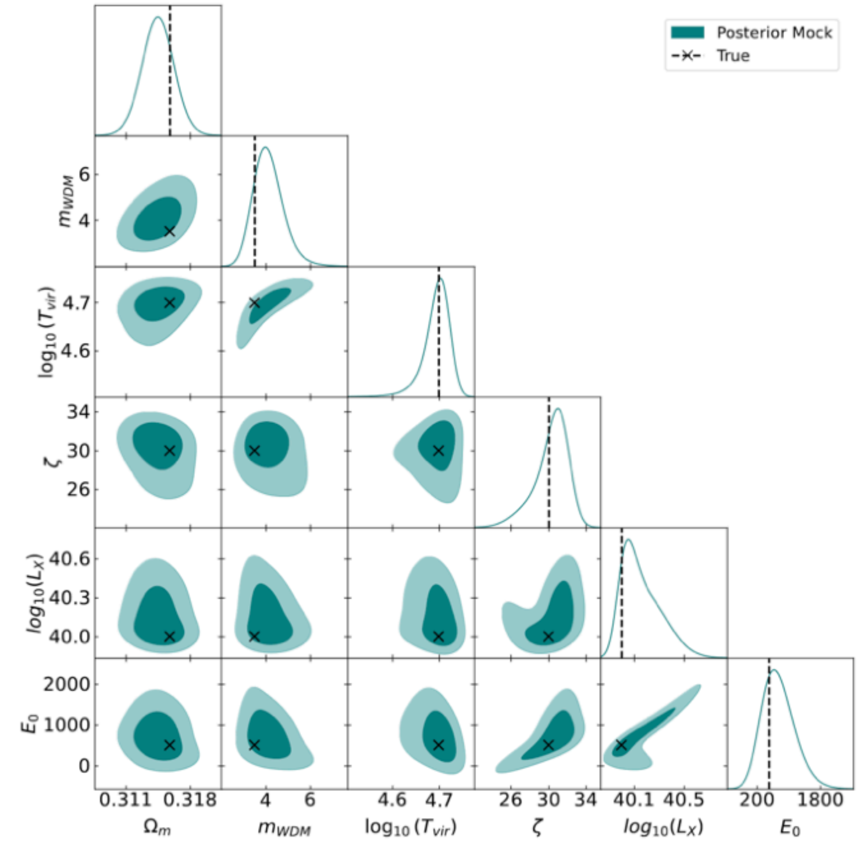


SBI with flows/cINN

based on BayesFlow

arXiv:2003.06281





Sim: Summary stays close to original

Mock: Heavy adjustment of summary vector

We profit from learned summary in presence of noise (more)!

Neutsch, Heneka, Brüggen (2022), arXiv:2201.07587

Schosser, Heneka, Plehn, arXiv:2401.04174

3) Simulation-based inference (SBI) for the SKA

Performance validation via:

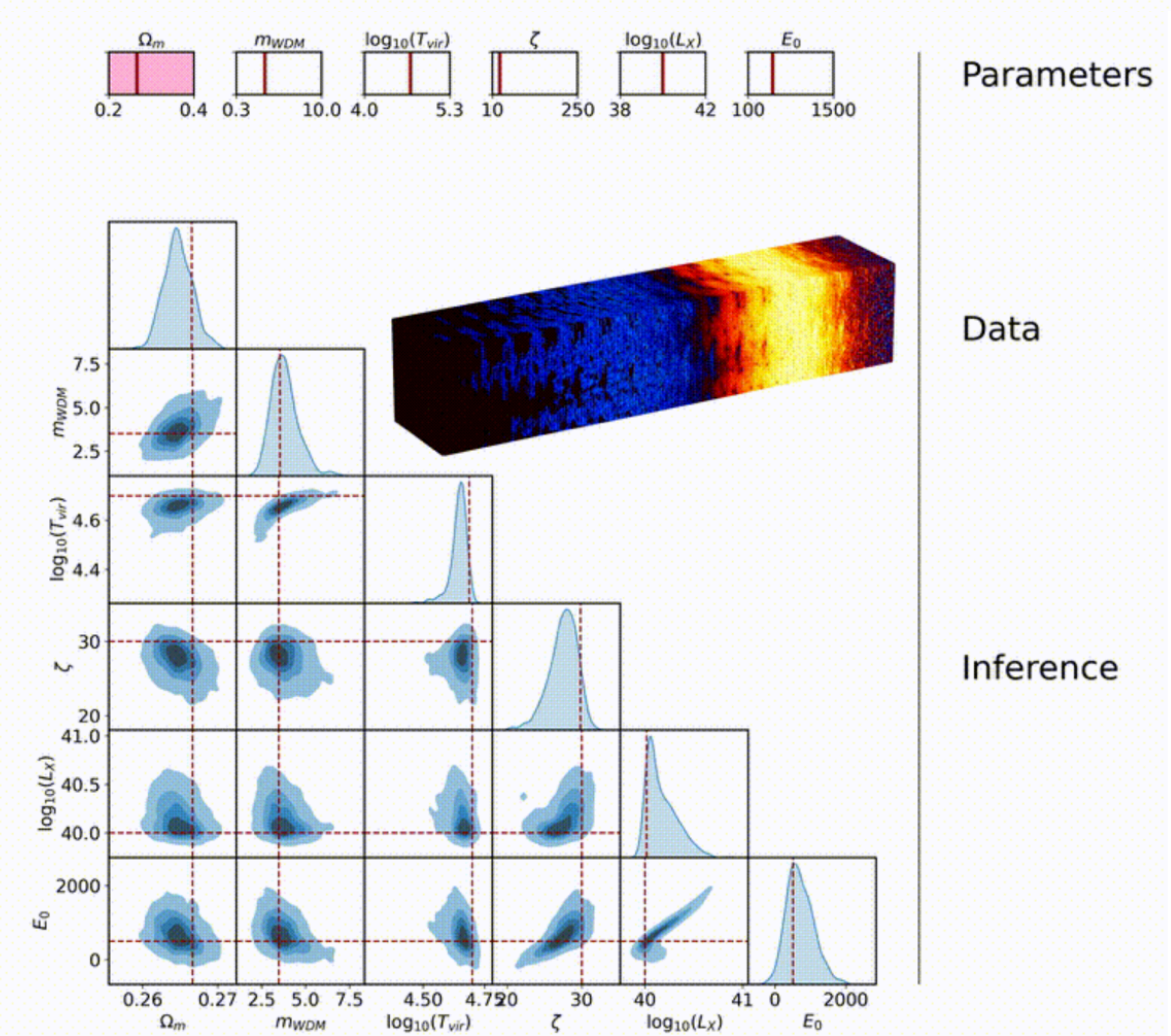
- Distribution of latent variables
- Simulation-based calibration
- Parameter recovery
- Mutual information

Trained SBI in action:

1 frame = 1 MCMC

astro-ML

21cm_pie Public



Credit: Benedikt Schosser

'Optimal, fast, and robust inference of reionization-era cosmology with the 21cmPIE-INN'

Schosser, Heneka, Plehn (2024), arXiv:2401.04174

3) Simulation-based inference (SBI) for the SKA

Performance validation via:

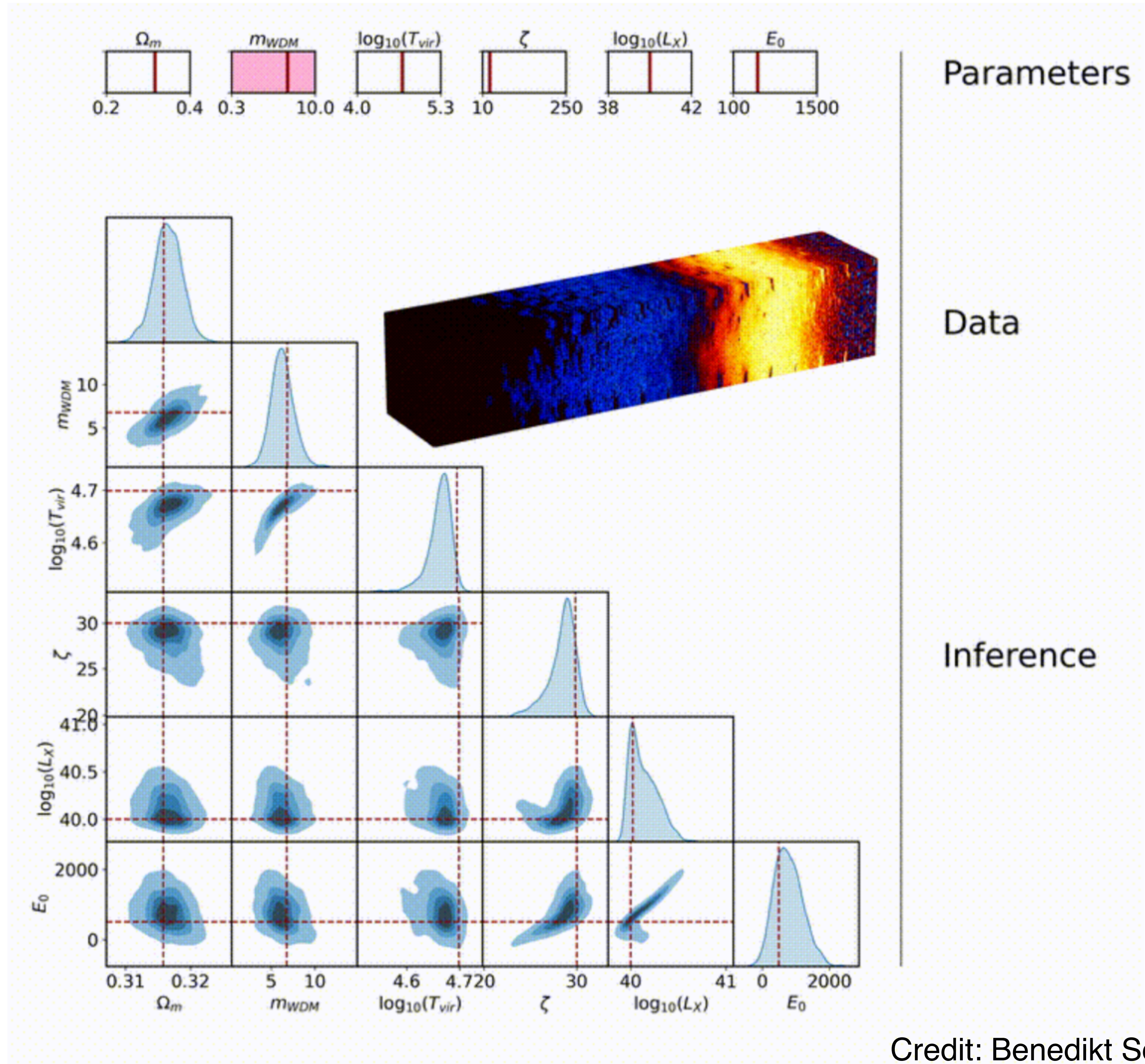
- Distribution of latent variables
- Simulation-based calibration
- Parameter recovery
- Mutual information

Trained SBI in action:

1 frame = 1 MCMC

astro-ML

21cm_pie Public



Credit: Benedikt Schosser

'Optimal, fast, and robust inference of reionization-era cosmology with the 21cmPIE-INN'

Schosser, Heneka, Plehn (2024), arXiv:2401.04174

3) Simulation-based inference (SBI) for the SKA

Performance validation via:

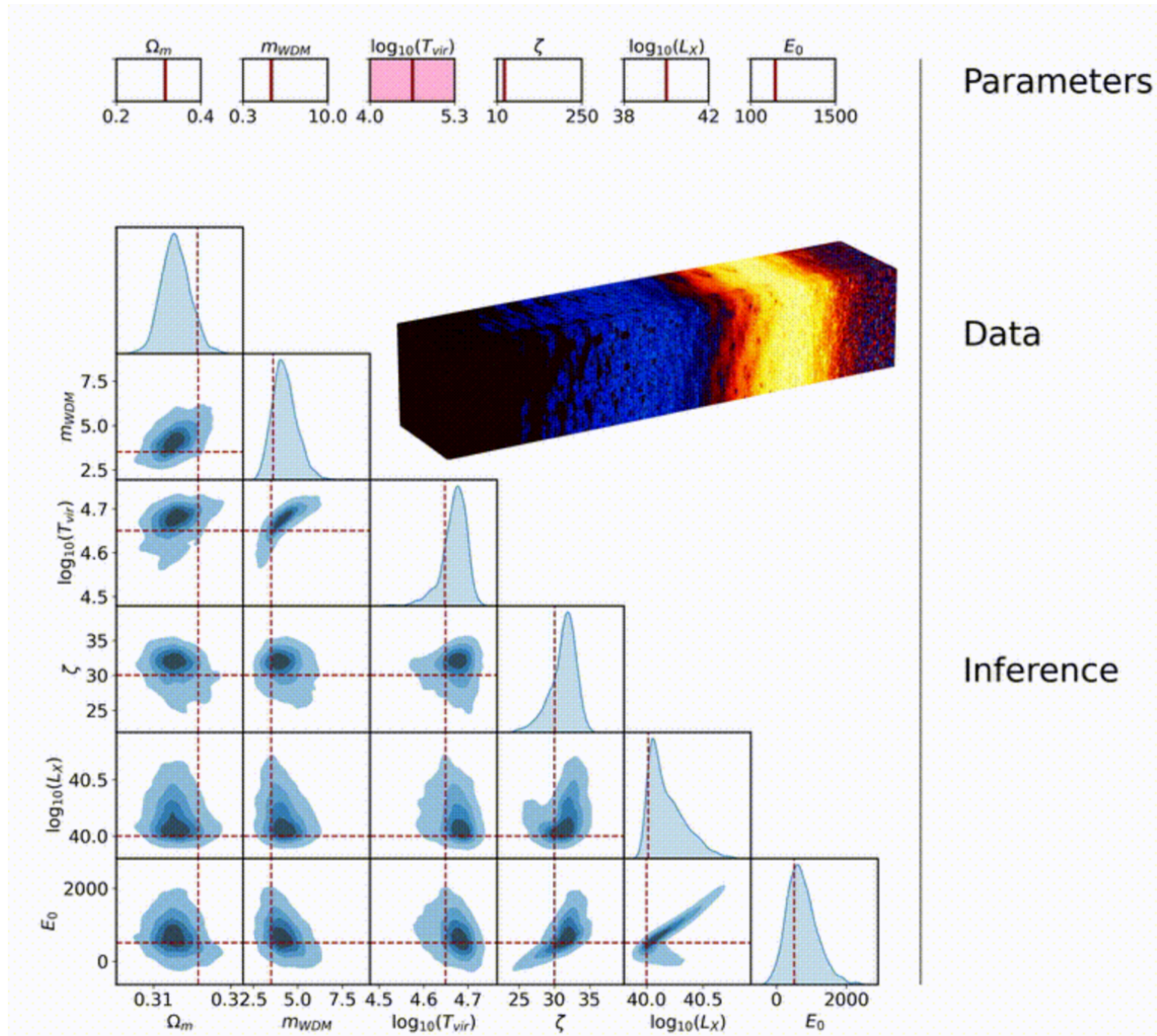
- Distribution of latent variables
- Simulation-based calibration
- Parameter recovery
- Mutual information

Trained SBI in action:

1 frame = 1 MCMC

astro-ML

21cm_pie Public



Credit: Benedikt Schosser

'Optimal, fast, and robust inference of reionization-era cosmology with the 21cmPIE-INN'

Schosser, Heneka, Plehn (2024), arXiv:2401.04174

3) Simulation-based inference (SBI) for the SKA

Performance validation via:

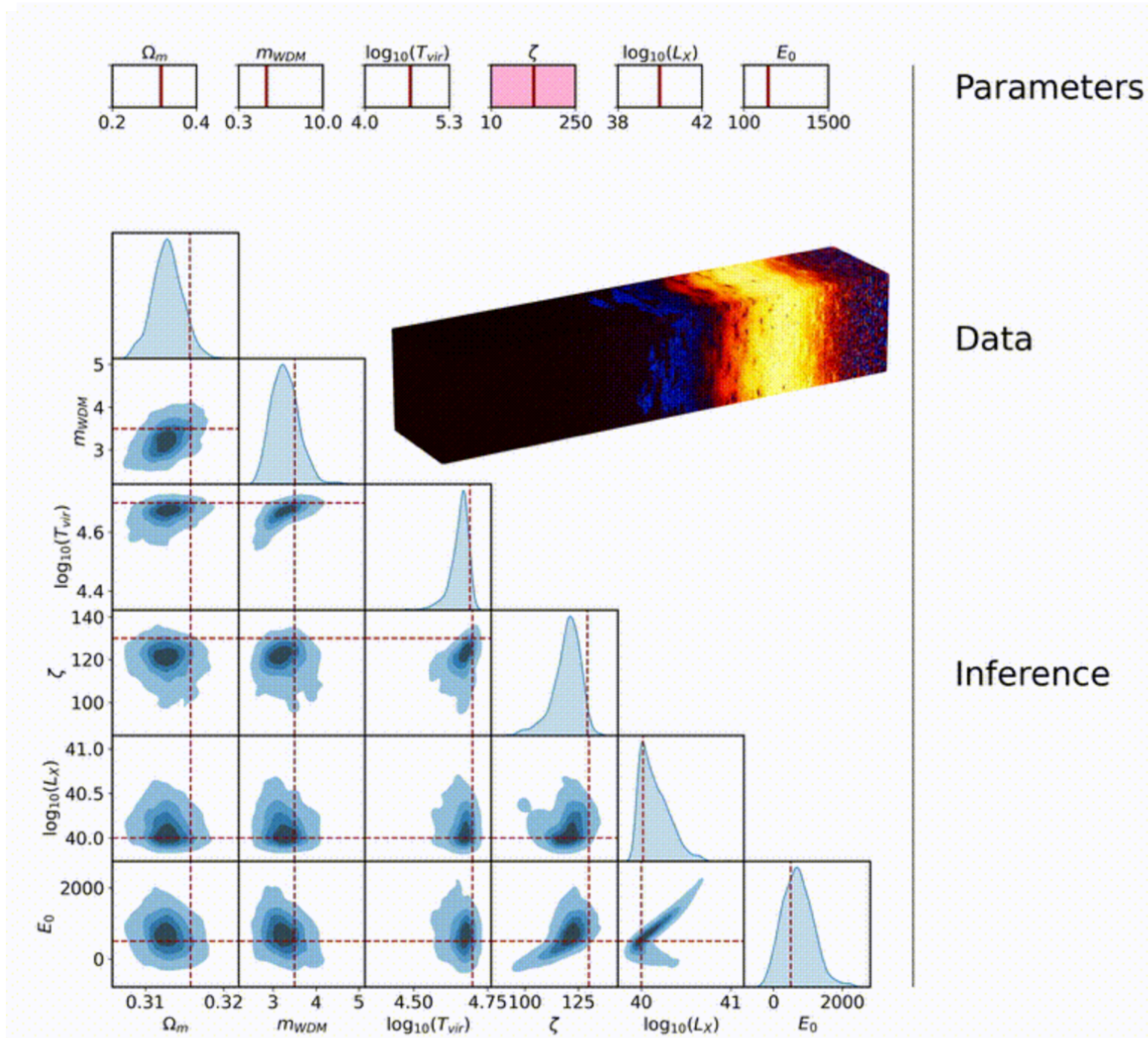
- Distribution of latent variables
- Simulation-based calibration
- Parameter recovery
- Mutual information

Trained SBI in action:

1 frame = 1 MCMC

astro-ML

21cm_pie Public



Credit: Benedikt Schosser

‘Optimal, fast, and robust inference of reionization-era cosmology with the 21cmPIE-INN'

Schosser, Heneka, Plehn (2024), arXiv:2401.04174

3) Simulation-based inference (SBI) for the SKA

Performance validation via:

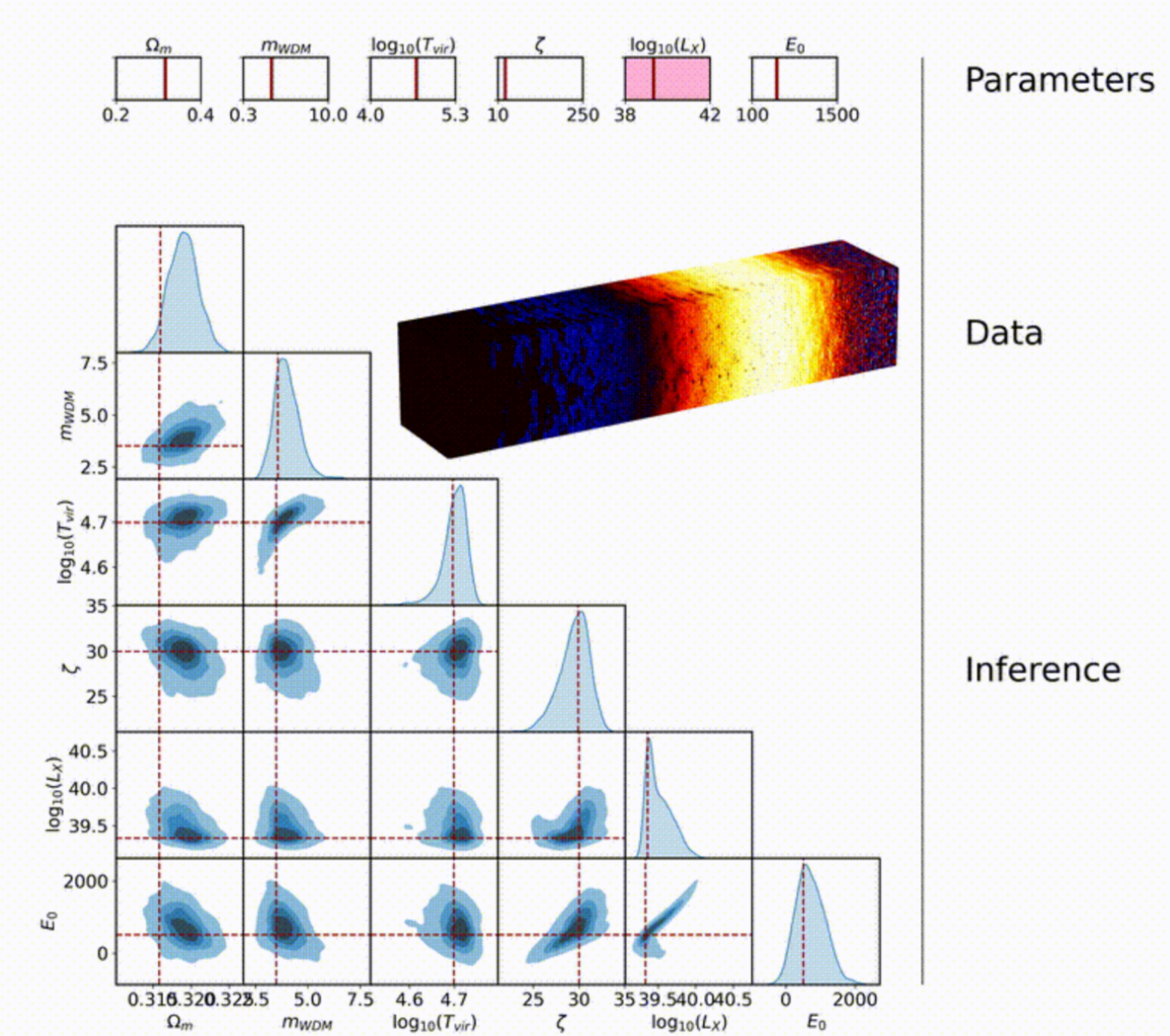
- Distribution of latent variables
- Simulation-based calibration
- Parameter recovery
- Mutual information

Trained SBI in action:

1 frame = 1 MCMC

astro-ML

21cm_pie Public



Credit: Benedikt Schosser

'Optimal, fast, and robust inference of reionization-era cosmology with the 21cmPIE-INN'

Schosser, Heneka, Plehn (2024), arXiv:2401.04174

3) Simulation-based inference (SBI) for the SKA

Performance validation via:

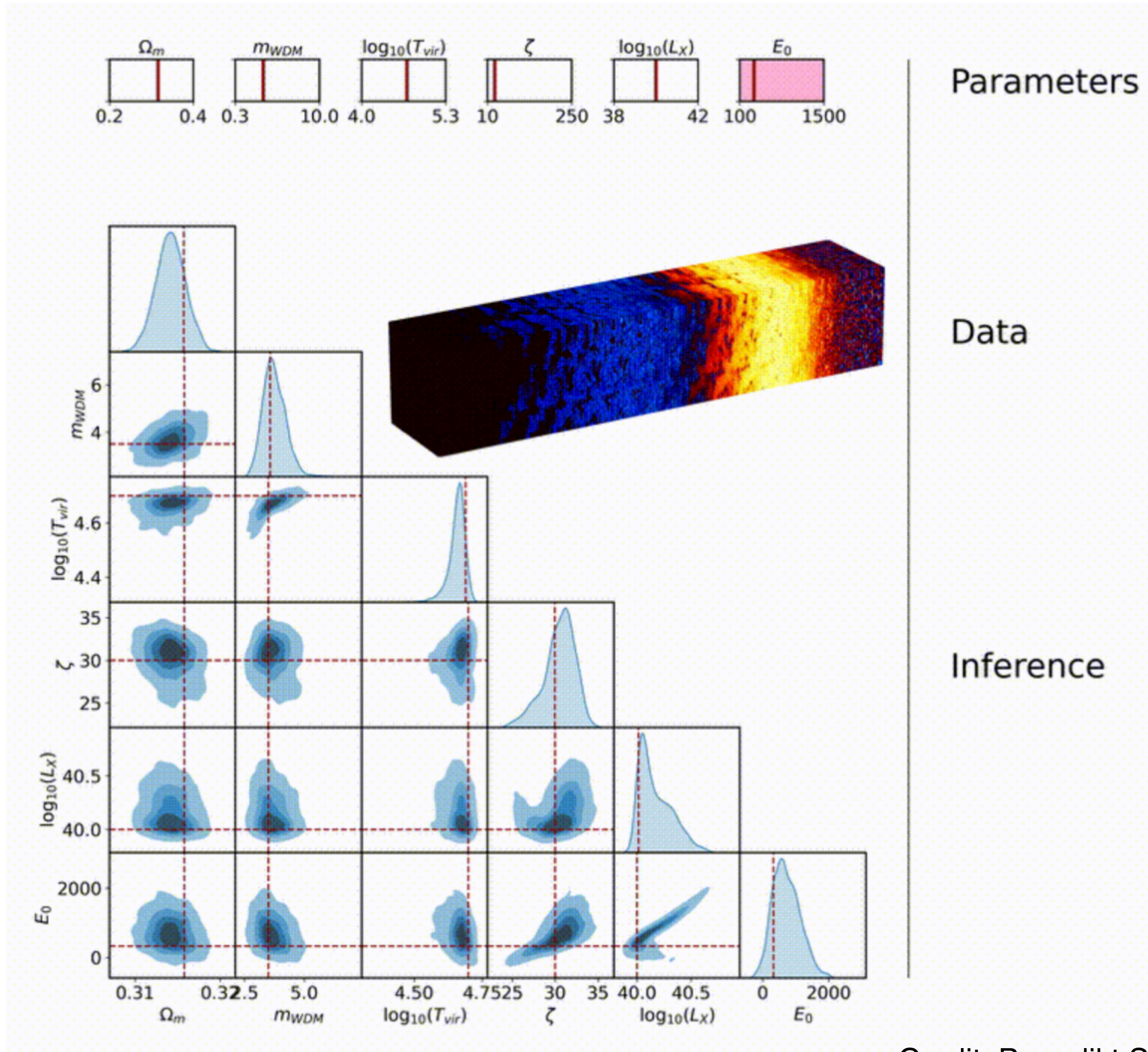
- Distribution of latent variables
- Simulation-based calibration
- Parameter recovery
- Mutual information

Trained SBI in action:

1 frame = 1 MCMC

astro-ML

21cm_pie Public

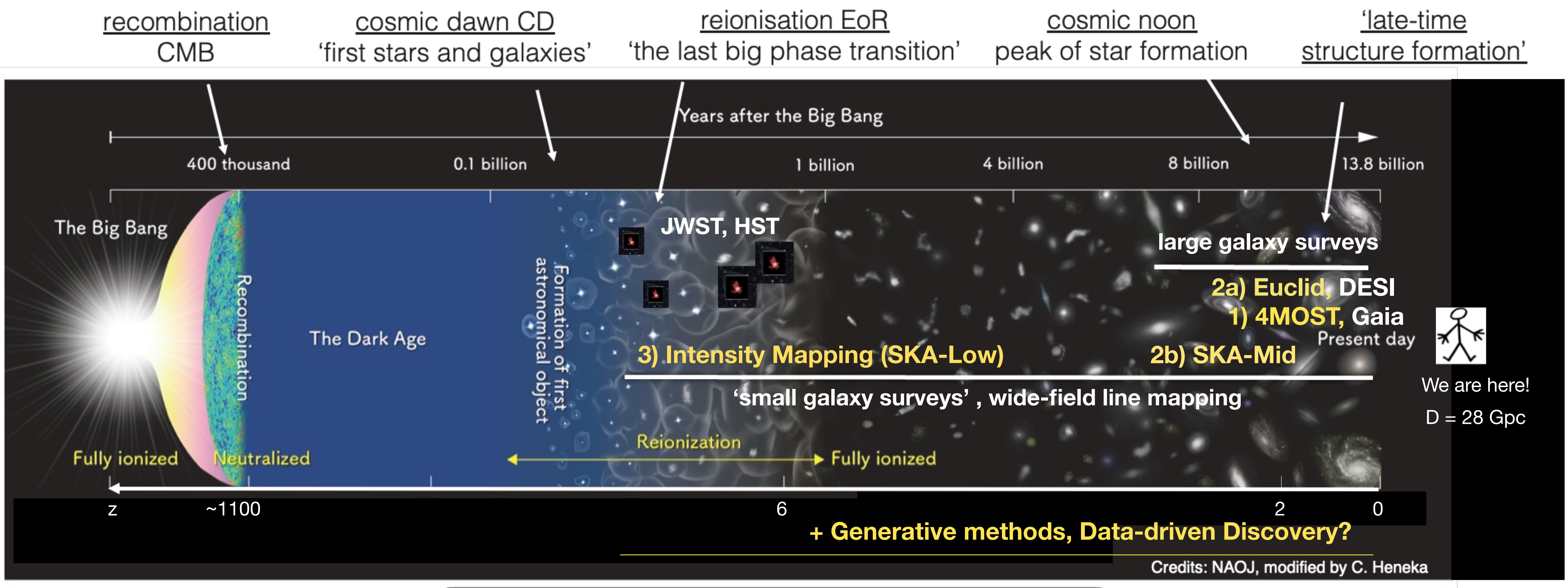


Credit: Benedikt Schosser

'Optimal, fast, and robust inference of reionization-era cosmology with the 21cmPIE-INN'

Schosser, Heneka, Plehn (2024), arXiv:2401.04174

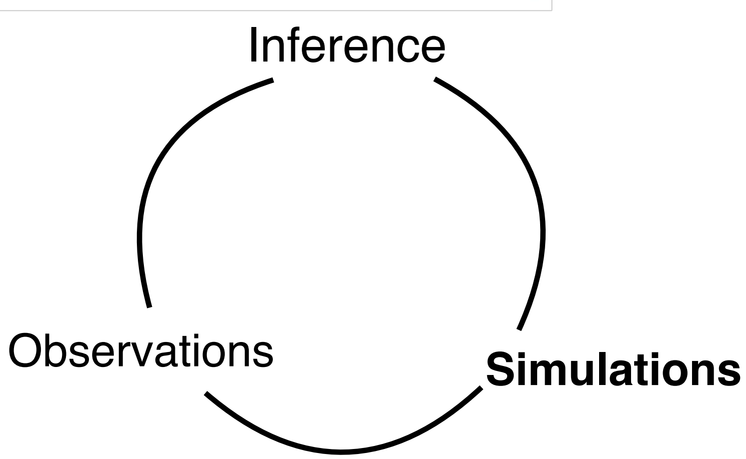
Research Highlights: Astronomical Data Science and Artificial Intelligence



Select Highlights

1) Classification
2) Source detection & characterisation
3) Simulation-based inference

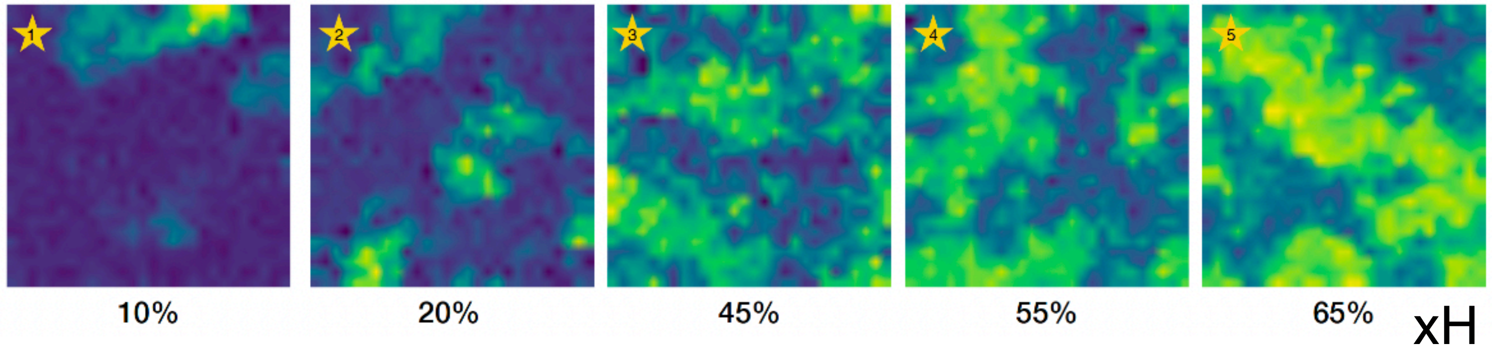
\+ **Generative methods, Data-driven Discovery**



+ Generative Modelling, Data-driven Discovery

Generative models:

Is there a fast way to emulate whole simulations?

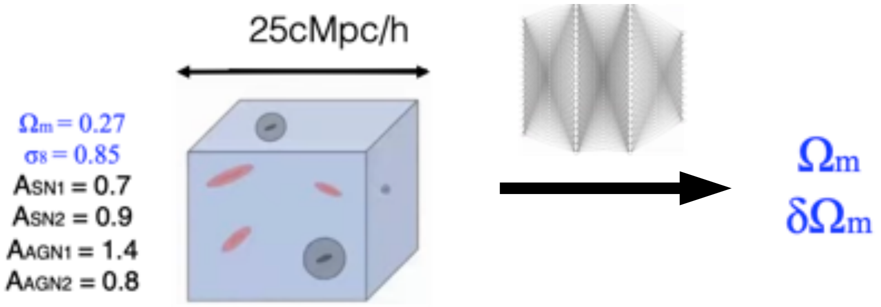


Diffusion models (Ho+20)

Trained on 21cmFAST
(Mesinger+12, Murray+20)

@Lara Alegre (Postdoc ITP)

+



YES!

~10% uncertainty

Connections on a high-dim manifold?

Villaescusa-Navarro (incl. Heneka)+22

Data-driven discovery:

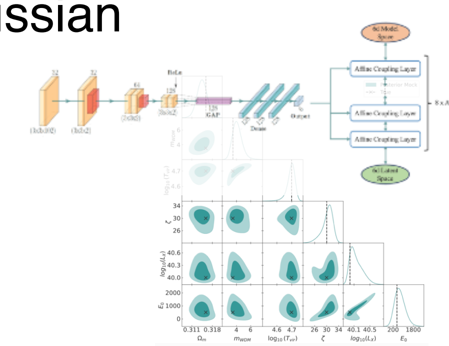
Can we measure Ωm only from one (random) galaxy?

… on the way to scientific discovery with ML/AI/Big Data and the SKAO !?

Summary: Where we stand

Goal: Understanding & discovery

Inference
SBI in 3D from non-Gaussian
tomographic data



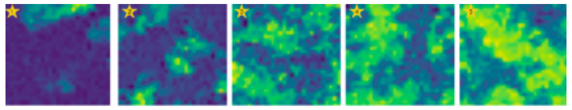
Inference
& Analysis

Detection & Characterisation
Unbiased measurements from
diverse sources (galaxies)

Data-driven **Discovery**

In astronomy & astrophysics

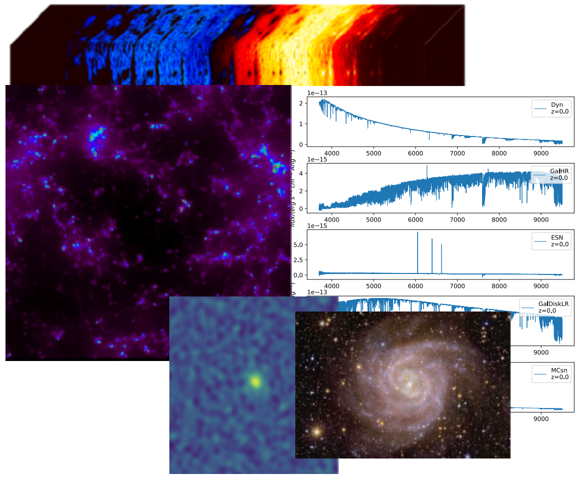
Simulation
Produce large range of
reionization topologies



Classification
Online classifier and triggering

Observations
& Data

Simulations
& Modelling



Select publications:
Neutsch, Heneka, Brüggen 22, arXiv:2201.07587
Schosser, Heneka, Plehn, arXiv:2401.04174
Hartley+ 23 (incl. Heneka), arXiv:2303.07943
Heneka 23, arXiv:2311.17553
Boucaud, Huertas-Company, Heneka+ 20, arXiv:1905.01324
Zhong, Napolitano, Heneka+24, arXiv:2311.04146

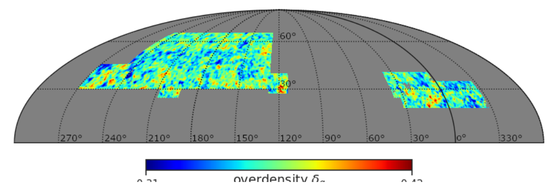
Next stop: Robust Foundational models

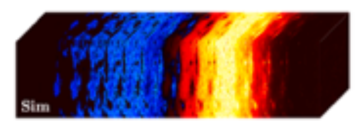
Generation & modelling

Data-Simulation Gap

Mock observations

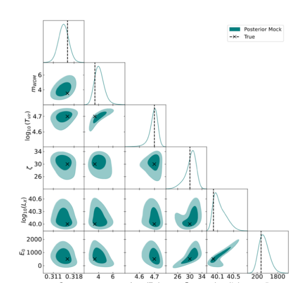
1) **Data, Mocks**





2) Transfer

3) Inference

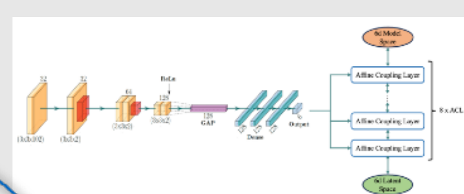


Interferometric observations

Maps / tomography

Simulation-based:

Optimal Summaries



Foundational ?

tuning, self-supervised
importance weighing

Stay tuned: Ayo Ore et al.

Currently:

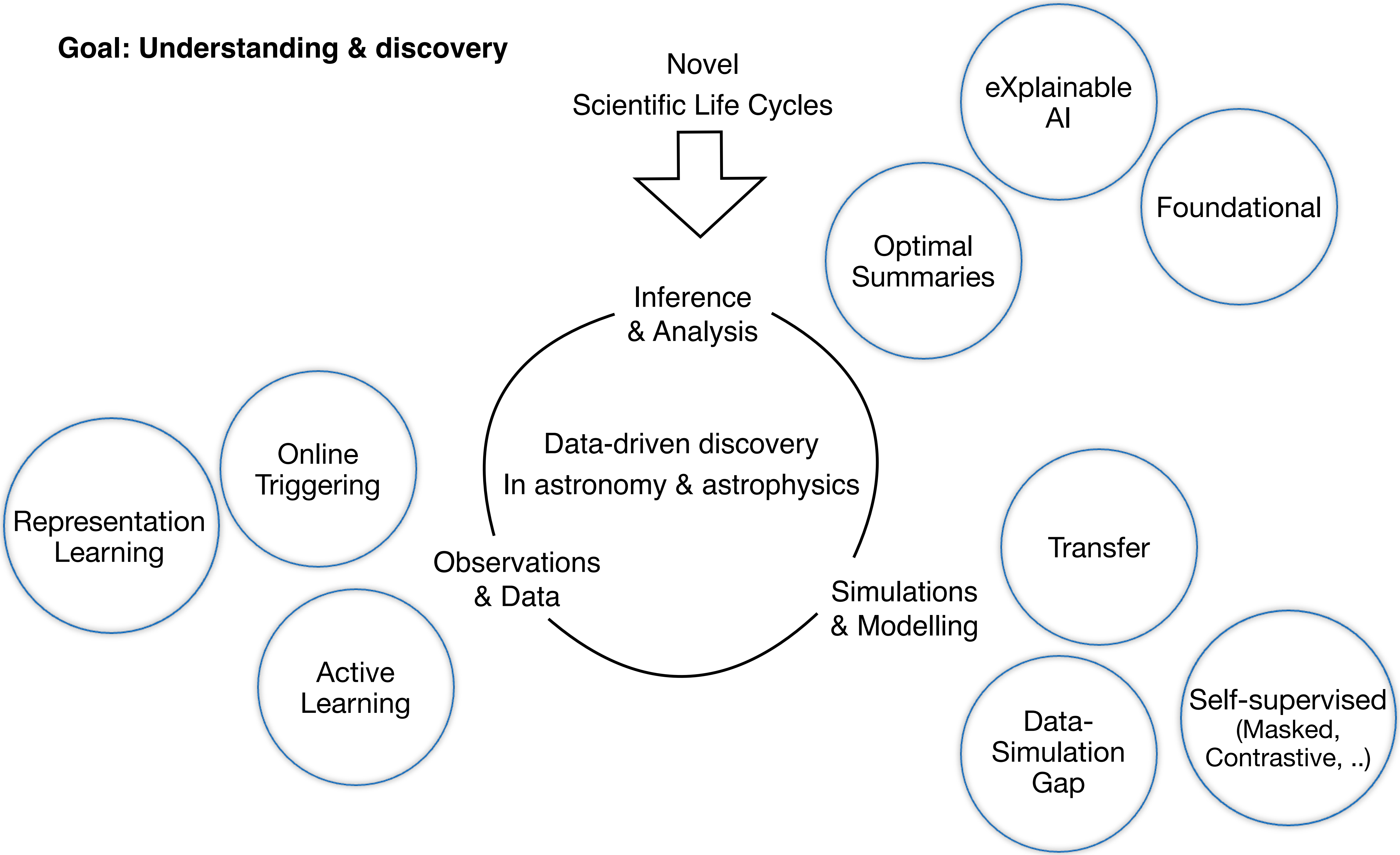
Comparison to ‘random mocks’
Derive summaries, such as C_ℓ^{gg}

+ jackknife
+ MCMC sampling

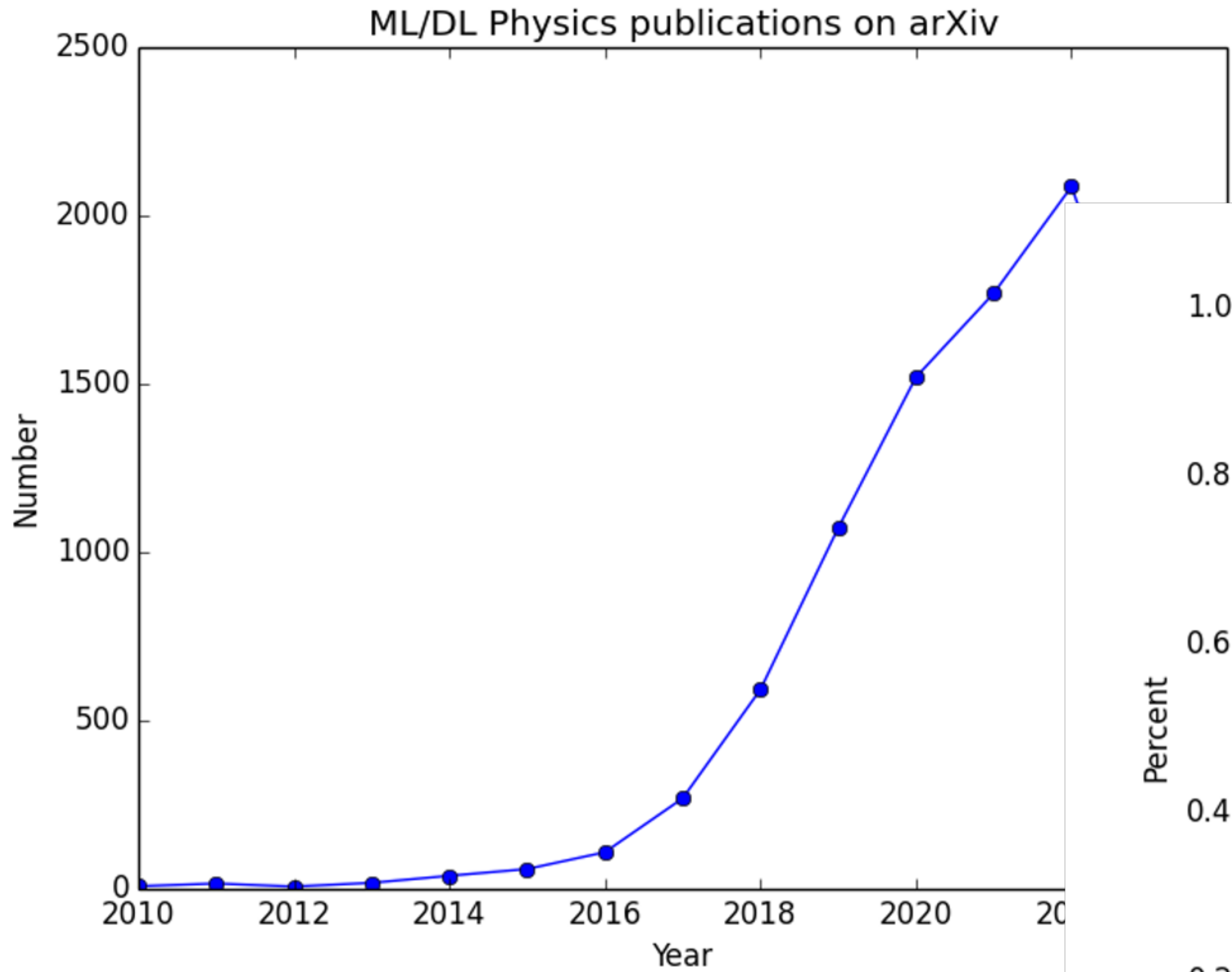
Next: Where we are going

‘What will bring our community forward’

Goal: Understanding & discovery

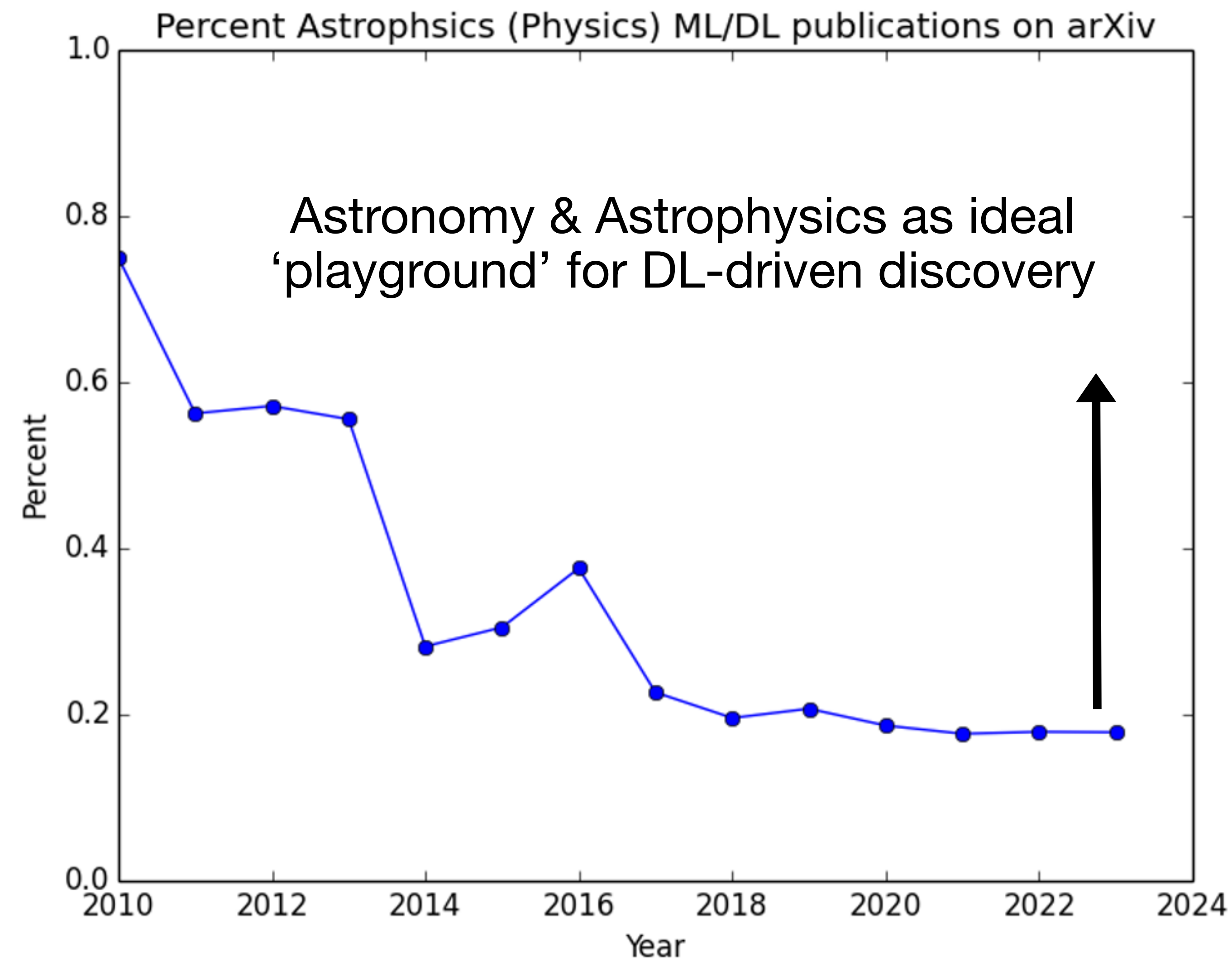


Summary and conclusions



Some advertisement:

We have ‘endless’ open-source public data



Thank you for your attention!

heneka@thphys.uni-heidelberg.de